Validitätsaspekte bei der Messung von Schreibkompetenzen

Dissertation
zur Erlangung des
akademischen Grades
Doctor rerum naturalium (Dr. rer. nat.)
im Fach Psychologie

eingereicht an der
Lebenswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von Thomas Canz M. A. M. A.

Präsident der Humboldt-Universität zu Berlin
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Lebenswissenschaftlichen Fakultät
Prof. Dr. Richard Lucius

Gutachter/Gutachterin:

1. Prof. Dr. Matthias Ziegler

2. Prof. Dr. Petra Stanat

3. Prof. Dr. Michael Becker-Mrotzek

Tag der Verteidigung: 19.10.2015

Auch getrennte Freunde mit süßen Banden zu knüpfen,
fand die gute Natur uns eine Sprache: die Schrift!
Sie führt Seelen zusammen, die fern aneinander gedenken,
führt den Seufzer herbei, der in den Lüften verhallt.

(Johann Gottfried von Herder, 1786)

Inhalt

Zusammenfassung

Schreibkompetenzen erweisen sich als zentral für den schulischen Erfolg und die kulturelle Teilhabe an der Gesellschaft, sie zu messen ist eine wichtige Aufgabe im Rahmen des Bildungsmonitorings. Diese Arbeit stellt eine nationale Bildungsstudie zur Erfassung von Schreibkompetenzen von Schülerinnen und Schülern am Ende der Sekundarstufe I für das Fach *Deutsch* vor. Das in dieser Studie angewandte Verfahren zur Schreibkompetenzmessung wird hinsichtlich drei ausgewählter Validitätsaspekte untersucht.

In der ersten Teilstudie wird geprüft, ob es sich bei *Schreibkompetenz* um ein textmusterunabhängiges Konstrukt handelt oder ob von textmusterspezifischen Schreibkompetenzen auszugehen ist. Diese Fragestellung wird auch für theoretisch angenommene Schreibkompetenzdimensionen, d. h. inhaltliche, stilistische und orthografisch-grammatische Schreibkompetenzen untersucht. Darüber hinaus wird auch die interne Struktur von *Schreibkompetenz*, die Beziehungen zwischen den genannten Schreibkompetenzdimensionen betrachtet. Die vorwiegend auf Modellvergleichen basierenden Analysen zeigen, dass es sich bei allgemeinen Schreibkompetenzen sowie bei inhaltlichen und stilistischen Schreibkompetenzen um textmusterspezifische Konstrukte handelt, bei der orthografisch-grammatischen Schreibkompetenz hingegen um ein textmusterunabhängiges Konstrukt. Für alle Textmuster zeigt sich eine zweidimensionale Struktur mit inhaltlich-stilistischen Schreibfähigkeiten als eine und orthografisch-grammatische Schreibfähigkeiten als zweite Dimension.

In der zweiten Teilstudie wird untersucht, inwiefern konstruktirrelevante Lesefähigkeiten bei der Messung von Schreibkompetenzen aufgrund der textuellen Präsentation der Aufgabeninstruktion miterfasst werden. Dabei werden die Instruktionstexte nach leseschwierigkeitsbestimmenden Merkmalen klassifiziert und der Einfluss der Lesekompetenz als Zusammenhangsstärke zwischen Lese- und Schreibkompetenz in Abhängigkeit von diesen Merkmalen in Form von Mehrebenenmoderatoranalysen bestimmt. Dabei zeigen sich statistisch bedeutsame, aber praktisch kaum relevante Effekte für zwei der schwierigkeitsbestimmenden Merkmale, die syntaktische Komplexität der Instruktionstexte und die Seltenheit der verwendeten Wörter.

In der dritten Teilstudie wird untersucht, ob eine von der sprachlichen Richtigkeit unabhängige Beurteilung inhaltlicher und stilistischer Schreibkompetenzen erfolgt oder ob hierbei Halo-Effekte zutage treten. In Anschlussanalysen wird geprüft, ob diese Halo-Effekte von der Art und Anzahl der Fehler, der Textlänge, der Textkomplexität und des Textmusters abhängen. Hierbei zeigen sich keine Urteilsverzerrungen bei der inhaltlichen, jedoch bei der stilistischen Bewertung. Diese Verzerrungen erweisen sich als größer bei syntaktisch komplexeren Texten sowie bei höherer Fehleranzahl und treten vor allem unter Vorliegen grammatischer Fehler und syntaktisch relevanter Zeichensetzungsfehler auf.

Abstract

Writing competencies are highly relevant for educational success and cultural participation in society; measuring these competencies is an important task within the scope of educational monitoring. This dissertation presents a national educational study assessing writing competencies in *German* of students at the end of lower secondary education. The underlying process of measuring writing competencies is investigated regarding three selected aspects of validity.

The first research study addresses the question, whether *writing competence* is a discourse mode independent construct or whether there are rather discourse mode dependent writing competencies. This question is also addressed with regards to the writing competence dimensions, i.e. contentual, stylistic and orthographic-grammatic writing competencies with an additional look on the relations between these dimensions. Analyses, which are predominantly based on comparison of IRT-models, reveal that writing competencies as well as the contentual and stylistic dimensions are discourse-mode-specific constructs, whereas the orthographic-grammatic writing competence is discourse mode independent.

The second research study raises the question to what extent – due to the fact that the writing task instructions are given textually – construct-irrelevant reading competencies are included when measuring *writing*. Therefore the instructions are classified with regards to aspects quantifying reading difficulty. Two-level moderator analyses are computed, modelling the correlation between *reading* and *writing competence* dependent on these aspects. Statistically significant but practically hardly relevant effects obtain for two of these aspects, i.e. the syntactic complexity of the instruction text and the (in)frequency of the used words.

The third research study investigates whether the evaluation of contentual and stylistic writing competencies takes place regardless of the orthographic and grammatic features of the evaluation underlying texts or whether halo effects occur. Further analyses examine possible rating shifts in dependence of error types, error amount, text length, text complexity and discourse type. The results reveal that stylistic, but not contentual rating shifts based on halo effects occur. These shifts are more pronounced in regard to syntactically more complex texts as well as higher error quantities and arise particularly under the presence of grammatical and syntactically relevant punctuation errors.

Ausführliches Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Dokumentenverzeichnis

1. Einleitung

1.1. Die Bedeutsamkeit von Schreibkompetenzen

„Lesen und Schreiben – Mein Schlüssel zur Welt“, so lautet eine aktuelle Alphabetisierungskampagne des Bundesministeriums für Bildung und Forschung (Bundesministerium für Bildung und Forschung, 2012; Hubertus, 2013) in Anlehnung an ein Wilhelm von Humboldt zugeschriebenes Zitat, der bereits vor 200 Jahren „Sprache [als] Schlüssel zur Welt“ (Borsche; 1989; Gipper, 1959; Schorer, 1959; Walter, 2003) beschrieb. Die heutige Anpassung dieses Slogans und Fokussierung der schriftsprachlichen Fähigkeiten *Lesen* und *Schreiben* unterstreichen die Bedeutung der Literalität in unserer modernen Gesellschaft des 21. Jahrhunderts.

Eine erfolgreiche gesellschaftliche Teilhabe lässt sich ohne schriftsprachliche Kenntnisse kaum noch vorstellen. Ein Großteil der beruflichen und privaten Kommunikation erfolgt aufgrund zeitlicher und örtlicher Distanz der Kommunikationspartner in Form schriftlicher Korrespondenz. Aufgrund der veränderten Lebensformen wird heute eine Vielzahl an Informationen, welche bis vor einigen Jahrzehnten noch mündlich tradiert wurden, in schriftlicher Form weitergegeben. Die gesellschaftlichen Veränderungen der letzten Jahrzehnte führte auch dazu, dass vor allem auch die Beherrschung produktiver schriftsprachlicher Fähigkeiten, die bis in die Mitte des letzten Jahrhunderts lediglich für Angehörige bestimmter sozialer Schichten und/oder Berufsbranchen eine alltagsrelevante Kompetenz darstellte, relevant für jedermann wurde (Arlt & Beelitz, 1970; Crotti & Osterwalder, 2008; Houston, 2011).

Auch aufgrund der zentralen Stellung, welche die Neuen Medien inzwischen einnehmen, verlagern sich einst prototypische Face-to-Face-Interaktionen immer stärker in den virtuellen Raum, womit vor allem nicht nur die schriftliche Rezeption, sondern auch die schriftliche Darstellung unerlässlich wird (Androutsopoulos, 2007; Fishman, Lunsford, McGregor & Otuteye, 2005; Schmitz, 2006).

Die Vermittlung schriftsprachlicher Fähigkeiten stellt daher ein fundamentales Ziel schulischer Bildung dar. Bis an das Ende der Sekundarstufe I, ein Zeitpunkt, welcher für einen Großteil der Schülerschaft das Ende der schulischen Laufbahn darstellt, sollten alle Schülerinnen und Schüler Kompetenzen in diesem Bereich erworben haben, die sie zu einer erfolgreichen Teilnahme am gesellschaftlichen Leben befähigen. Der Vermittlung von

Schreib- und Textproduktionsfähigkeiten kommt hierbei im Muttersprachunterricht eine zentrale Bedeutung zu (Becker-Mrotzek & Böttcher, 2014; Gätje, 2013; M. Fix, 2006; Harsch, Neumann, Lehmann & Schröder, 2007; Merz-Grötsch, 2010; NAEP, 2003).

Aufgrund einer dominanten Fokussierung auf die Vermittlung schriftsprachlicher Fähigkeiten im Muttersprachunterricht (Nutz, 2006; Schöler, 2006; Schuster, 1998) ist die Fähigkeit zu schreiben auch für den schulischen Erfolg selbst unmittelbar relevant. Da der schulische Erfolg maßgeblich weitere berufliche und soziale Entwicklungen bedingt, kommt der Fähigkeit zu schreiben eine in doppelter Weise zentrale Stellung zu.

1.2. Die Messung von Schreibkompetenzen

Aufgrund der schulischen, beruflichen und sozialen Relevanz von Schreibfähigkeiten besteht der Bedarf, Schreibkompetenzen reliabel und valide erfassen zu können. Im schulischen Rahmen zeigte sich, dass klassische Aufsatzbenotungen sehr stark subjektiven Urteilen unterliegen, teilweise durch schreibkompetenzirrelevante Faktoren beeinflusst sind und somit nicht diesen Gütekriterien genügen (Ingenkamp, 1971; Schröter, 1971). Daher schien es unerlässlich, die Zensurgebung um reliable und valide Tests zu ergänzen (Schöler, 2006; Tent, 1998) (vgl. Kapitel 2.3.1.).

In Deutschland wurden zu Beginn des neuen Jahrtausends länderübergreifende Bildungsstandards eingeführt, welche „allgemeine Bildungsziele [definieren und festlegen] (…), die Kompetenzen die Kinder und Jugendlichen bis zu einer bestimmten Jahrgangsstufe erworben haben sollten“ (Klieme et al., 2007, S. 19). Diese Zielvorgaben schufen einen einheitlichen Referenzrahmen für die Entwicklung von Instrumenten zur systematischen empirischen Überprüfung dieser Ziele (vgl. Kapitel 2.1.4.).

Für den Kompetenzbereich *Schreiben* stand eine solche Überprüfung vor besonderen Herausforderungen, weshalb große nationale und internationale Schulleistungsstudien wie PISA oder die IQB-Ländervergleiche diesen Kompetenzbereich bisher aussparten. Neben einem deutlich erhöhten Zeit- und Kostenaufwand gegenüber Kompetenzerfassungen in anderen Bereichen ist die Messung von Schreibkompetenzen mit inhaltlichen und methodischen Problemen und Fragen verbunden, die sich nicht nur im Rahmen der Überprüfung der Bildungsstandards stellen, sondern die das Schreibassessment im Allgemeinen betreffen.

So lassen sich Schreibkompetenzen nur unzureichend in einzelne Teilfähigkeiten zerlegen, die sich in isolierten Teilaufgaben (Items) erfassen lassen. Die Erfassung von Schreibkompetenzen, welche Personen befähigen, Texte zu produzieren, kann nur anhand der Beurteilung von Texten erfolgen, die als Resultate von Schreibhandlungen vorliegen. Texte sind jedoch hochkomplexe sprachliche Gebilde, die im Hinblick auf eine Vielzahl sprachlicher und inhaltlicher Aspekte betrachtet und bewertet werden können. Eine Schwierigkeit besteht nun darin, festzulegen, welche Aspekte die relevanten, d. h. textgütebestimmenden sind und in welcher Form sie am geeignetsten zu erfassen sind, beispielsweise anhand feingliedriger kriterialer Detailurteile oder anhand abgestufter ganzheitlicher Urteile (weitere Ausführungen hierzu unter Kapitel 2.3.2.).

Eine weitere Schwierigkeit stellt die Reliabilität der Bewertung dar, d. h. die Frage, auf welche Weise und inwiefern eine zuverlässige bewerterunabhängige Beurteilung der Aspekte gewährleistet werden kann (vgl. Kapitel 2.3.2 & 3.5.).

Nicht zuletzt ist die Messung und Überprüfung von Schreibkompetenzen auch mit einer Reihe von Fragen der Validität dieser Messung bzw. der Ergebnisse (vgl. Kapitel 5) konfrontiert, d. h. Fragen danach, wie das Konstrukt *Schreibkompetenz* geeignet erfasst und modelliert werden sollte oder ob und inwiefern eine vollständige und exklusive, von konstruktexternen Faktoren befreite Erfassung des Konstrukts gelingt (vgl. Kapitel 5–9).

1.3. Ziel und Struktur dieser Arbeit

Die vorliegende Arbeit verfolgt das Ziel, eine Studie zur Messung von Schreibkompetenzen von Schülerinnen und Schülern am Ende der Sekundarstufe I vorzustellen und das angewandte Verfahren im Hinblick auf zentrale Validitätsaspekte zu überprüfen.

Empirische Basis für diese Darstellungen und Untersuchungen bildet die Studie zur Normierung von Aufgaben zur empirischen Überprüfung des Erreichens der Bildungsstandards für den Mittleren Schulabschluss (MSA) im Kompetenzbereich *Schreiben* des Faches *Deutsch*, welche im Auftrag der Länder durch das Institut zur Qualitätsentwicklung im Bildungswesen (IQB) durchgeführt wurde.

In Kapitel 2 wird zunächst der schulpraktische, wissenschaftliche und bildungspolitische Hintergrund dargestellt, welcher zur relevanten Auffassung von *Schreibkompetenz* sowie zur Erfassung des Konstrukts im Rahmen von Large-Scale-Bildungsstudien beigetragen hat.

Darüber hinaus erfolgt eine Diskussion und Klärung der Begriffe *Kompetenz* (im Allgemeinen) und *Schreibkompetenz* (im Spezifischen). Abschließend wird ein Überblick über zentrale, für diese Arbeit relevante Aspekte der empirischen Schreibleistungsforschung gegeben.

Anschließend wird in Kapitel 3 die Durchführung der Normierungsstudie detailliert, von der Aufgabenentwicklung bis zur Genese und Verabschiedung der Kompetenzstufenmodelle, beschrieben. Besonderes Gewicht fällt hierbei auf die Kodierung, d. h. auf die Bewertung der Schreibaufgaben (Beschreibung und Diskussion der verwendeten Schemata; Reliabilitätsbetrachtungen) und die Beschreibung des Standard-Setting-Verfahrens, ein Verfahren zur Ermittlung der Stufengrenzen für das zu generierende Kompetenzstufenmodell. Grund für die ausführliche Deskription des Standard-Setting-Verfahrens ist die Tatsache, dass dieses Verfahren hochgradig kompetenzbereichsspezifisch ist und im Rahmen der Normierungsstudie in dieser Form erstmalig im deutschsprachigen Raum eingesetzt wurde.

In Kapitel 4 werden die Hauptbefunde der Normierungsstudie deskriptiv dargestellt und anschließend mit Blick auf Ergebnisse bisheriger Schulleistungsstudien eingeordnet und diskutiert. Die Ergebnisdarstellungen konzentrieren sich hierbei auf ermittelte Kompetenzstufen und Bewertungen anhand eines der beiden eingesetzten Auswertungssysteme (*holistisches Kodierschema*, vgl. Kapitel 3.3.). Neben der Betrachtung der Gesamtschülerschaft erfolgt auch eine vergleichende Betrachtung nach Klassenstufe, Geschlecht, Schulabschluss, Schulform und Sprachhintergrund.

Vor dem konzeptuellen und empirisch-praktischen Hintergrund, der in den Kapiteln 2 bis 4 etabliert wurde, stellt das Folgekapitel den messtheoretischen Rahmen für die noch folgenden Forschungsarbeiten dar. In Kapitel 5 wird der Validitätsbegriff vorgestellt und diskutiert; darüber hinaus werden verschiedene Validitätsaspekte erläutert und zu den folgenden Forschungsfragestellungen in Bezug gesetzt.

In den Kapiteln 6 bis 8 werden die drei zentralen Forschungsarbeiten dieser Dissertation vorgestellt. Aufgrund der dominant messtheoretischen und empirischen Rahmung dieser Arbeiten findet sich eine ausführliche themenspezifische Darstellung von Theorie und Forschungsstand jeweils in den einzelnen Kapiteln und nicht in einem übergreifenden vorgeschalteten Theoriekapitel. Generell ist diese Arbeit strukturell so konzipiert, dass es dem Leser mit wenigen Ausnahmen (beispielsweise beim Verweis auf die gemeinsame

Datengrundlage oder die eingesetzten Kodierschemata)[1] ermöglicht sein soll, einzelne Hauptkapitel isoliert zu rezipieren.

Kapitel 6 geht hierbei der Hauptfragestellung nach, inwiefern es sich bei Schreibkompetenzen von Schülerinnen und Schülern am Ende der Sekundarstufe I um ein textmusterunabhängiges Konstrukt (*Schreibkompetenz*) oder um textmusterspezifische Schreibkompetenzen (*narrative Schreibkompetenz*, *argumentierende Schreibkompetenz*, *informierende Schreibkompetenz*) handelt. Darüber hinaus wird geprüft, wie sich die Frage nach der Textmusterunabhängigkeit bzw. Textmusterspezifität für inhaltliche, stilistische und orthografisch-grammatische Schreibfähigkeiten beantworten lässt; dabei wird auch auf das Verhältnis dieser drei Dimensionen zueinander eingegangen.

Kapitel 7 verfolgt die Fragestellung, ob und in welchem Maße bei der Erfassung von *Schreibkompetenz* aufgrund des Einsatzes von Schreibaufgaben, in welchen die Instruktion und einige inhaltliche und stilistische Vorgaben textuell dargeboten werden, konstruktexterne Lesefähigkeiten miterfasst werden.

Auch in Kapitel 8 wird die Erfassung möglicher nichtbeabsichtigter Aspekte beleuchtet. Hier wird untersucht, ob bei der Bewertung rein inhaltlicher bzw. rein stilistischer Textqualitätsaspekte Verstöße gegen die sprachliche Richtigkeit, d. h. orthografische und grammatische Fehler, zu Urteilsverzerrungen führen. In Zusatzanalysen wird geprüft, ob diese Verzerrungen von bestimmten Fehler- und/oder Textmerkmalen abhängen.

In Kapitel 9 erfolgt schließlich eine Zusammenfassung, Gesamtschau und übergreifende Diskussion der Ergebnisse sowie ein Ausblick auf mögliche zukünftige Forschungsfragen im Rahmen des Themenfelds.

1.4. Formelle und begriffliche Festlegungen dieser Arbeit

Aufgrund der Interdisziplinarität der vorliegenden Arbeit, der zugrunde liegenden Forschung und somit auch der antizipierten Leserschaft sei an dieser Stelle auf einige formelle und begriffliche Konventionen, Festlegungen und Besonderheiten hingewiesen, die ggf. nicht in allen Disziplinen Anwendung finden bzw. nicht allen Lesern vertraut und/oder transparent sind.

[1] Kapitelübergreifende Verweise sind an den entsprechenden Stellen expliziert. Häufig dienen diese (lediglich) zur vertiefenden (Detail-)Information.

Häufig wird in der vorliegenden Arbeit auf Begriffe, Variablen, Konstrukte etc. in Form von Bezeichnungen, die als Namen fungieren, referiert. Diese Namen sind durch Kursivdruck gekennzeichnet. Solche Ausdrücke werden nicht in Deklination und Groß-/Kleinschreibung angepasst. Dies gilt nicht für gleichlautende Ausdrücke, insofern sie nicht als Namen verwendet werden.[2]

Als Dezimaltrennzeichen in numerischen Ausdrücken werden (gemäß internationalen wissenschaftlichen Publikationsstandards und entgegen den gültigen Rechtschreibregeln des Deutschen) Punkte statt Kommata verwendet.

Ausführungen zu unterschiedlichen Textmustern fokussieren in der vorliegenden Arbeit drei dieser Muster, für die folgender begrifflicher Gebrauch gilt: Für das argumentierende Textmuster werden (je nach syntaktischer Funktion) die Ausdrücke *argumentieren*, *argumentativ*, *argumentierend*, *Argumentation* und *(das) Argumentieren* verwendet. Für das informierende Textmuster sind nach Einschätzung des Autors nicht alle analogen Ausdrücke geeignet, da sie mit einer anderen Bedeutung belegt sind (*Information*, *informativ*), daher werden hier lediglich die Ausdrücke *informieren*, *informierend*, *(das) Informieren* verwendet. Für das narrative Textmuster werden prädominant die Begriffe *narrativ* und *Narration* gebraucht; in Kontexten, in welchen sich jedoch auf andere Autoren und deren Sprachgebrauch bezogen wird oder eine syntaktische Funktion realisiert werden soll, wofür im Deutschen kein stammidentisches Wort existiert, werden bisweilen auch die Ausdrücke *erzählen* oder *(das) Erzählen* verwendet; *narrativ* und *erzählend* (und entsprechende Wortformen anderer Wortarten) werden hier somit synonym verwendet. Aufgrund der nicht-parallelen Wortbildungsmöglichkeiten der Ausdrücke für die drei verschiedenen Textmuster kommt es oftmals zu sprachlich nicht parallelen Wortgegenüberstellungen, die Nennformen gehören hier oftmals sogar verschiedenen Wortklassen an, bspw. *argumentieren – informieren – narrativ.*

Darüber hinaus wird auf der Ebene theoretischer Schreibkompetenzaspekte und potentieller Schreibkompetenzdimensionen zwischen den Aspekten/Dimensionen *Inhalt*, *Stil* und *sprachliche Richtigkeit* unterschieden. Um syntaktisch umständliche Formulierungen in entsprechenden Teilen zu vermeiden, schien es an den entsprechenden Stellen notwendig, auch adjektivisch auf diese Aspekte/Dimensionen referieren zu können. Während dies für

[2] Beispiele zur Illustration: „Zur Diskussion des Begriffes *Schreiben* …“ (aber: „Zur Diskussion des Kompetenz-begriffs …“), „Für das Textmuster *argumentieren* zeigte sich …“ (aber: „Für das argumentierende Textmuster zeigte sich …“

Inhalt (*inhaltlich*) und *Stil* (*stilistisch*) wortstammident und sprachlich transparent möglich ist, bietet das Deutsche kein entsprechendes Adjektiv für *sprachliche Richtigkeit* (in der Bedeutung: *die sprachliche Richtigkeit betreffend*); aus diesem Grund wird in solchen Kontexten das Adjektivkompositum *orthografisch-grammatisch*, welches auf die betroffenen Teilaspekte der Dimension verweist, verwendet.

Des Weiteren wird im Rahmen dieser Arbeit prädominant von *Schreibkompetenzen* die Rede sein. Die Verwendung des Plurals beruht zum einen auf empirischen Evidenzen zur Textmusterspezifität von *Schreibkompetenz* und somit von Schreibkompetenzen, zum anderen auf theoretischen Überlegungen sowie empirischen Befunden zu verschiedenen Schreibkompetenzdimensionen. Von *Schreibkompetenz* (Singular) ist zumeist im Kontext theoretischer gesamtkonstruktbezogener Fragestellungen sowie im Rahmen begrifflicher Erläuterungen die Rede. Da die differenzierungsrelevanten Gründe erst im Fortgang der Arbeit transparent werden, sei der Leser an dieser Stelle auf diese begriffliche Besonderheit hingewiesen.

2. Hintergrund

Im Folgenden wird in Kapitel 2.1. ein kurzer Abriss über die verschiedenen Strömungen gegeben, welche zur heutigen Art und Weise der Schreibkompetenzerfassung im Bildungswesen beigetragen haben, darunter die Textlinguistik, die schulische Praxis und Fachdidaktik, die Psycholinguistik und Kognitionspsychologie sowie die Empirische Bildungsforschung und die Bildungspolitik. Im Anschluss erfolgt in Kapitel 2.2. eine Klärung des für diese Arbeit zentralen Begriffes *Schreibkompetenz*. In Kapitel 2.3. werden der Forschungsstand, Probleme, Fragen und bisherige zentrale und für diese Arbeit relevante Befunde der empirischen Schreibleistungsforschung dargestellt. Daran angebunden werden das weitere Vorgehen in dieser Arbeit sowie die Fragestellungen der Forschungsteilstudien erläutert.

2.1. Historischer Abriss

2.1.1. Textlinguistik

Um die modernen fachdidaktischen Ansätze innerhalb der Schreibforschung zu verstehen, ist es wichtig, die Entwicklungen im Bereich der Textlinguistik, welche sich mit den Produkten des Schreibens wissenschaftlich auseinandersetzt, zu kennen. Die Textlinguistik ist neben der Rhetorik (vgl. Kapitel 2.1.2.) und kognitionswissenschaftlichen Ansätzen (vgl. Kapitel 2.1.3.) eine der wichtigsten Einflussquellen der Schreibforschung (Baurmann & Weingarten, 1995; Sieber, 2006)

In der Linguistik herrschte bis in die 1960er Jahre die syntaktische Forschung vor. Der Satz galt bis dahin als die linguistische Untersuchungseinheit (Gansel & Jürgens, 2009). Erst durch die zunehmende bewusste Detektion von grammatischen Phänomenen, welche sich nicht unter Bezug auf Einzelsätze erklären ließen, wurden Sätze verstärkt in ihrem Kontext betrachtet und die Textlinguistik entstand (Brinker, 1985; P. Hartmann, 1971).

Zunächst dominierte hierbei ein strukturell-grammatischer Ansatz, der die bisherigen linguistischen Beschreibungsebenen (Phonem – Morphem – Wort – Satzglied – Satz) um die textuelle Ebene ergänzte (W. Heinemann & Viehweger, 1991). Dabei wurde davon ausgegangen, dass „Texte strukturelle Einheiten vom gleichen Typ wie Sätze sind, nur

umfangreicher“ (Vater, 1992, S. 20). Eines der Ziele dieses strukturell-grammatischen Ansatzes war es, Textkohärenz anhand von koreferenten Ausdrücken (bspw. Nomen und zugehörige Pronomen) zu erklären (Gansel & Jürgens, 2008; Harweg, 1968). Allerdings wurde bald klar, dass sich einige Kohärenzphänomene nicht mit diesem Ansatz auffangen ließen, und semantische Textbetrachtungen rückten in den Fokus der Textlinguistik (Brinker, 1985).

Bereits vor Entstehung der Textlinguistik wurden im Rahmen der Prager Schule (Mathesius, 1929) semantische Textgliederungsprozesse betrachtet. Im Rahmen des Konzepts einer funktionalen Satzperspektive stand auch hier der Satz als dominante Untersuchungseinheit im Fokus, jedoch wurden Satzbildungsprinzipien auf satzübergreifende, d. h. textuelle und außersprachliche Faktoren, zurückgeführt. Das zentrale Konzept im Prager Strukturalismus war die Thema-Rhema-Struktur von Sätzen. Unter *Thema* wird dasjenige verstanden, worüber etwas ausgesagt wird; es ist prototypisch, das Bekannte, die bereits im Text vorerwähnte oder durch den Gesamtkontext bereits gegebene Information. Das *Rhema* entspricht dem Neuen, dem Ausgesagten, dem Mitteilenswerten (Ludger Hoffmann, 2000; Brinker, 1985). Ein Text wird demnach als eine Reihe von Sätzen verstanden, welche nach bestimmten thematisch-rhematischen Abfolgeprinzipien strukturiert sind (Daneš, 1970). Ludger Hoffmann (2000) entwickelte diesbezüglich ein Modell der Themenentfaltung in Texten, welches sich an der Thema-Rhema-Strukturierung orientiert; in dessen Rahmen ist zentral, „dass über Themen rhematische Informationen angehäuft und systematisch ins Wissen integriert werden“ (S. 352).

Ebenfalls auf semantischen Kriterien beruhen propositionale Textauffassungen, welche auch aufgrund ihrer kognitionspsychologischen Fundierung prominent wurden (Bransford & Franks, 1971; Bransford, Barclay & Franks, 1972; Ratcliff & McKoon, 1976; Weisberg, 1986). Eine *Proposition* ist eine Aussage aus logischem Subjekt (= das, worüber etwas ausgesagt wird) und logischem Prädikat (= das, was ausgesagt wird) und ist die zentrale satzsemantische Kategorie (Kintsch, 1982; Polenz, 1988). Im Rahmen eines Textes werden Propositionen zu- und miteinander in Beziehung gesetzt. Es erfolgt der Aufbau eines propositionalen Netzwerkes, welches benachbarte Propositionen miteinander verknüpft und letztendlich die komplexe Bedeutungsstruktur des gesamten Textes widerspiegelt (Anderson, 1976; Kintsch, 1974). Diese Netze werden hierbei jedoch nicht willkürlich aufgebaut, vielmehr erfolgt der Aufbau strukturiert durch Bündelung von einzelnen Sinneinheiten zu

größeren, komplexen Sinneinheiten (Makrostrukturen); diese Makrostrukturen selbst stehen in einer hierarchischen Struktur zueinander (van Dijk, 1980).

Nach der pragmatischen Wende in der Linguistik, die dominant von der Sprechakttheorie von J. L. Austin (1972) und J. R. Searle (1969) eingeführt wurde und ihre Wirkung in der 1970ern entfaltete, rückten handlungs- und kommunikationstheoretische Aspekte in den Vordergrund des Interesses. Texte wurden im Hinblick auf ihre kommunikative Funktion (bspw. Informations-, Appell-, Obligations-, Kontakt- und Deklarationsfunktion; vgl. Brinker, 1985) betrachtet und beschrieben. Das Primat des (syntaktischen) Satzes wurde abgelöst durch ein Primat der kommunikativen Funktion einer Sprachhandlung bzw. deren Absicht und Zweck (*Illokution*) (U. Fix, 2008a). So untersuchten etwa Motsch und Viehweger (1981) im Rahmen dieses Ansatzes die Funktionen aller Einzelaussagen (*Illokutionen*) in Texten und versuchten diese in Illokutionshierarchien abzubilden und Texte entsprechend zu klassifizieren.

Inzwischen werden die drei oben genannten Ansätze der strukturell-grammatischen, der semantischen und der kommunikativ-pragmatischen Textbetrachtung integriert. Während in den ersten beiden der Blick vom Satz zum Text geht und die Textbetrachtung somit als Bottom-Up-Prozess zu verstehen ist, wird im pragmatischen Ansatz im Top-Down-Prozess der Blick vom Textganzen zu den Textteilen gerichtet (Viehweger, 1983). In der integrativen Betrachtung wird ein Text als „in sich kohärente Einheit der sprachlichen Kommunikation mit einer erkennbaren kommunikativen Funktion und einer in spezifischen Weise organisierten Struktur“ (Gansel & Jürgens, 2009, S. 51) betrachtet, Textfunktion und Textstruktur werden zueinander in Beziehung gesetzt (U. Fix, 2008a; Sandig, 2006); oftmals wird dieser Ansatz daher auch als *funktionale* Textbetrachtung kategorisiert (Baumann, 1992; Gansel & Jürgens, 2009; Welke, 1993).

Die Integration der verschiedenen Textbetrachtungsebenen im Rahmen des heutigen Textverständnisses weist bereits auf die Komplexität des Untersuchungsgegenstands *Text* hin, welche sich auch später in den Textbeurteilungsinstrumenten widerspiegelt (vgl. Kapitel 2.3.2., 3.3. & 3.5.).

2.1.2. Schulisches Schreiben in der Praxis und Fachdidaktik in Deutschland

Bis ins 18. Jahrhundert hinein wurde Schulbildung nur der gesellschaftlichen Elite zuteil. Das schulische Schreiben orientierte sich hierbei an der Rhetorik, das Verfassen von Texten wurde nach den damaligen Vorstellungen als Vorstufe zum mündlichen Vortragen angesehen. Rhetorische Ideale waren das Verwenden von sprachlichen Figuren und Tropen, was dem damaligen Zeitgeist entsprach, Schreiben als Kunst anzusehen, und in einer Zeit lange vor einer wissenschaftlichen Betrachtung von Texten (U. Abraham, 1996, 2014; Ludwig, 2006).

Gegen 1770 wurde das Primat der Rhetorik von der Stilistik abgelöst, oftmals wird von diesem Zeitpunkt als Etablierung des Schulaufsatzes gesprochen (Rupp, 1986; Ludwig, 1988). Mit dieser Wende entwickelten sich auch die ersten Auseinandersetzungen mit den Fragen, was schriftlicher Ausdruck ist und wie dieser umgesetzt werden sollte. Der Anspruch an die Schreibenden bestand nun nicht mehr darin, sich originell, expressiv und ästhetisch auszudrücken, vielmehr wurden „kognitive Klarheit und Nachvollziehbarkeit (…) zum neuen Richtziel des Schreibunterrichts“ (U. Abraham, 2014, S. 6).

Im 19. Jahrhundert bildete sich in Folge der stilistischen Textbetrachtung eine Art Formenlehre oder Gattungsdidaktik (U. Abraham 1996, 2014; Rupp, 1986). Zudem sorgte die allgemeine Schulpflicht und damit der Anspruch, Schriftlichkeit der gesamten Bevölkerung zu vermitteln, für eine Spaltung der Unterrichtsformen schulischen Schreibens. Im Vergleich zu dem stilistisch orientierten Unterricht an den Gymnasien, erfolgte an den Volksschulen eine Konzentration auf die Vermittlung von basalen Schreibfähigkeiten, so standen grammatische Übungen oder das Kopieren und Imitieren von (Ultra-)Kurztexten (bspw. Quittungen) im Zentrum dieses Unterrichts (Ludwig, 1988; Seifert, 1835; Schießl, 1889).[3]

Im ausgehenden 19. Jahrhundert wurde die an Darstellungsformen orientierte Lehrweise von Aufsätzen im Rahmen des Volksschulunterrichts scharf kritisiert. Im Rahmen der damaligen naturalistischen Strömungen wurde nun vielmehr eine Vorstellung vertreten, dass die Schreiber am besten ohne formale Vorgaben unter größtmöglicher Freiheit „von selbst zum natürlichen Ausdruck [fänden]. (…) Unterstellt wird dabei, dass das Texteschreiben gleichsam naturwüchsig aus dem Ausdruckswillen junger Schreiber/-innen hervorquillt“ (U. Abraham 2014, S. 13); aus dieser Mentalität heraus entstand der freie Aufsatz der Reformpädagogik (Ludwig, 1988, Necknig, 2007).

[3] Dieser Unterschied im Schreibunterricht zwischen Volksschulen und Gymnasien hielt bis in die Mitte des 20. Jahrhundert an.

Bereits in den 1920ern kam es jedoch wieder zu einer Kehrtwende: Man rückte ab von einer Fokussierung auf Inhalte und die Stilbildung wurde ins Zentrum des Aufsatzunterrichts gestellt (Ludwig, 1988; Ritter, 2008). Es etablierte sich in Folge die Lehre des *sprachschaffenden* oder *sprachgestaltenden Aufsatz*, welche bis in die 1970er dominierte. Erneut wurde sich an Darstellungsformen orientiert, die basal in persönliche, subjektbezogene (z. B. Erzählungen oder Erörterungen) und sachliche, objektbezogene (z. B. Berichte oder Beschreibungen) unterschieden wurden (Ludwig, 1988). Becker-Mrotzek und Böttcher (2014) kritisieren, dass das Schreiben im Rahmen des sprachgestaltenden Ansatzes somit lediglich zu einem reinen Nachgestalten wurde; Ludwig (2003) konstatiert, dass dadurch „allmählich der deutsche Aufsatzunterricht geworden [ist], der sich auf die Vermittlung von gerade mal fünf Aufsatzarten und deren Einübung beschränkt." (S. 240).

Nutz (2006) beanstandet an der klassischen Form der Aufsatzlehre, dass „der traditionelle ‚Aufsatz' und seine Differenzierung in ‚Aufsatzarten' oder ‚Stilformen' eine sehr künstliche Einengung der Möglichkeiten schulischer Textproduktionen war und ist", und verweist auf die „Kluft zwischen schulischen Aufsatzarten und den Textsorten des öffentlichen Sprachgebrauchs" (S. 925). Engelen (1974) stellt hierbei einen zentralen Kritikpunkt heraus: „Der Hauptakzent lag dabei auf Aufbau und Stil. Ein eigentlicher Adressat fehlt" (Engelen, 1974, S. 243). Dies änderte sich mit der pragmatischen Wende der 1970er (vgl. Kapitel 2.1.1.). Durch die Fokussierung auf den Kontext, die kommunikative Situation, in welcher Sprachhandlungen stattfinden, wurde die Orientierung am Rezipienten, an Anlässen und Absichten zentral, das Schreiben für den Leser rückte in den Mittelpunkt (U. Abraham, 2014; Boettcher, Firges, Sitta & Tymister; Sauter & Pschibul, 1977).

In den kommenden beiden Jahrzehnten gab es einige weitere Strömungen in der Fachdidaktik, so etwa das *kreative Schreiben*, welches als Mittel der subjektiven Expression angesehen wurde und dem leserorientierten Schreiben ein schreiberorientiertes Schreiben gegenüberstellte (Spinner, 1980; 1993); in der schulischen Praxis sind diese Strömungen jedoch kaum angekommen (Merz-Grötsch, 2001). In der Praxis gewinnt vielmehr das literaturorientierte und interpretative Schreiben (bspw. Textinterpretationen) zunehmend stärkeres Gewicht (U. Abraham, 2014; Becker-Mrotzek & Böttcher, 2014). Becker-Mrotzek und Böttcher kritisieren diese Entwicklungen wie folgt: „Damit hatte der Deutschunterricht eine wichtige Funktion eingebüßt, (…) nämlich die Ausbildung einer fundierten Schreibkompetenz" (S. 73).

Diese Funktion der Vermittlung von Schreibkompetenz wurde im Laufe der 1990er Jahre (wieder) zentral. Schreibprozesse standen nun im Fokus des Interesses. Prozesse der Ideenfindung und Textplanung rückten ins Blickfeld, ebenso wie die Textüberarbeitung. Praktisch sorgte dies für eine stärkere Rückmeldekultur, wobei der Schreiber Leserreaktionen als Feedback erhielt (U. Abraham, 2014; Ludwig, 2006). Bei der Betrachtung dieser Schreibprozesse wurde das Augenmerk darauf gelenkt, was Schreibende *können*, der Begriff der Schreibkompetenz wurde zentral (M. Fix, 2006).

2.1.3. Kognitionspsychologische und psycholinguistische Forschung

In der psycholinguistischen und kognitionspsychologischen Forschung wurde das Schreiben eher stiefmütterlich behandelt. Dies mag nicht zuletzt an den experimentellen Methoden dieser Disziplinen liegen, worunter prädominant Reaktionszeitstudien, Augenbewegungsmessungen (eye-tracking) und seit den 1990er Jahren auch neurowissenschaftliche Verfahren wie EEG- und fMRT-Messungen fallen, Techniken somit, welche Informationen über die *on-line* stattfindenden Prozesse liefern (Rickheit, Herrmann & Deutsch, 2003; Rickheit, Sichelschmidt & Strohner, 2004). Diesen Messverfahren verschließt sich das Schreiben aufgrund der geplanteren und zeitlich ausgedehnteren Prozesse, die eine Erfassung der während der Sprachverarbeitung stattfindenden Prozesse in Echtzeit einschränken, sowie der parallelen motorischen Aktivität, welche Störeffekte und Interferenzen bei der Anwendung der oben genannten Messmethoden verursacht.

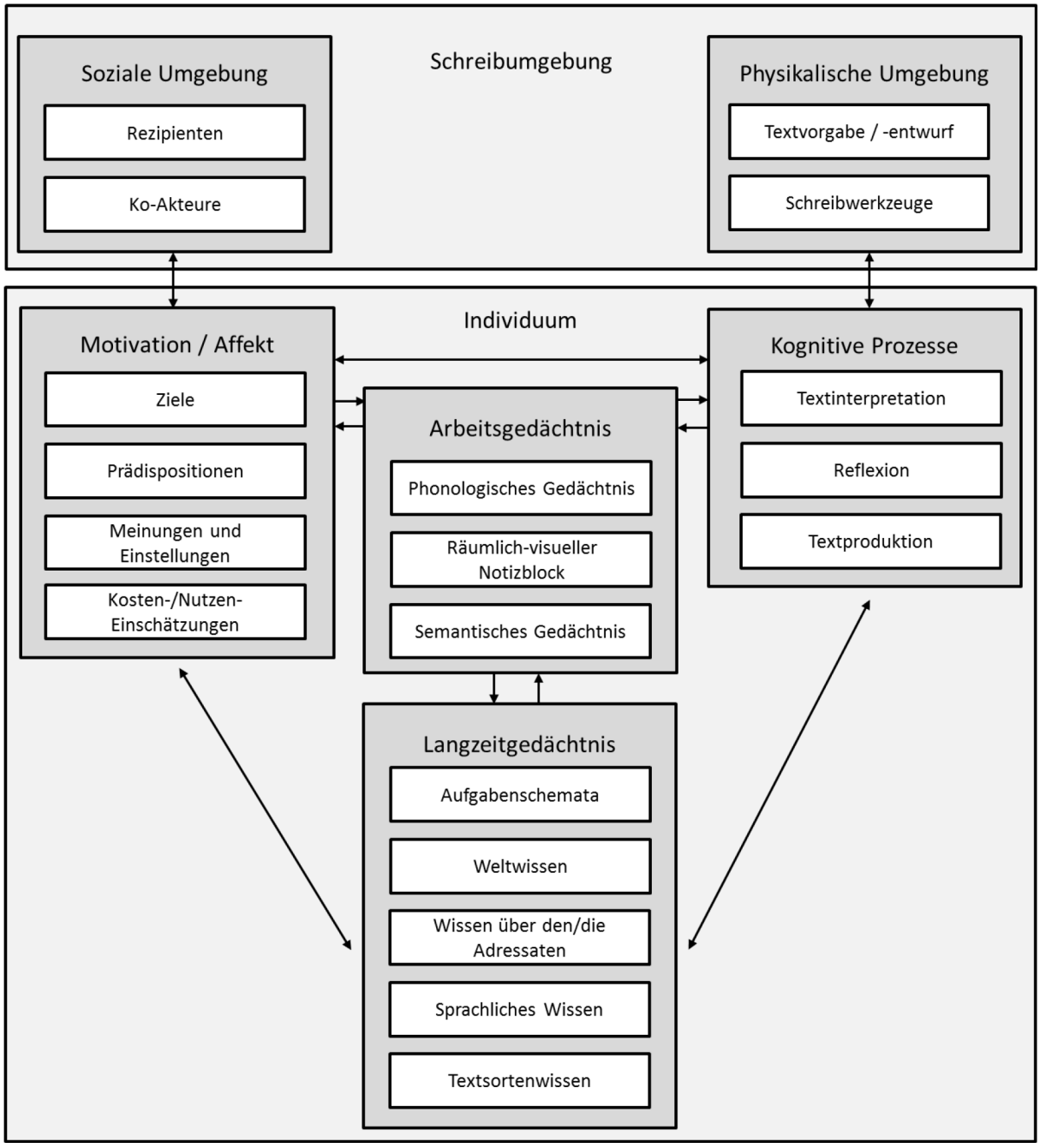
Dennoch führte die Schreibforschung in diesen Disziplinen zu einem bis heute einflussreichen und breit akzeptierten kognitiven Modell des Schreibens, welches auch die Prozessorientierung des modernen fachdidaktischen Ansatzes (vgl. Kapitel 2.1.2.) forcierte. Dieses Modell geht auf Hayes und Flower (1980) zurück und wurde in den Folgejahren um einige Aspekte ergänzt (Hayes, 1996; vgl. Abbildung 2.1.3.1). Das Modell wird hierbei vier zentralen Aspekten gerecht: Situierung in einem Kontext, Prozessualität des Schreibens, Einbeziehung motivationaler Zustände sowie Beachtung von aufgabenspezifischen Aspekten. (Folgende Ausführungen richten sich nach Flower & Hayes, 1981; Hayes & Flower, 1980; Hayes, 1996 und sind als Erläuterungen zu Abbildung 2.1.3.1 zu verstehen.)

So wird gemäß dem Modell der Schreibprozess in einem Kontext verortet, externe Merkmale wie zur Verfügung stehende materielle Ressourcen sowie die Leserschaft und mögliche Koautoren und andere Kollaborateure werden als mögliche fördernde bzw. hemmende Faktoren miteinbezogen.

Des Weiteren werden die unterschiedlichen Phasen des Schreibens (inklusive Planungs- und Revisionsprozesse) berücksichtigt. Der kognitive Kern der Schreibkompetenz besteht aus dem Wechselspiel zwischen dem Arbeitsgedächtnis (und dessen Instanzen), Inhalten des Langzeitgedächtnisses, worunter Aufgaben- und Textsortenwissen sowie Sprach- und Weltwissen fallen, jedoch auch der Grad der Kenntnis der Rezipienten, deren Vorwissen, Erwartungen und Ansprüche, und beim Schreiben beteiligte kognitive Prozesse wie das Produzieren oder Interpretieren des Textes bzw. Teile des Textes.

Auch motivationale und affektive Eigenschaften des Schreibers stehen in ständigem wechselseitigem Austausch mit den kognitiven Instanzen. So ist der Schreibprozess direkt und indirekt (über die kognitiven Ressourcen) von motivationalen Aspekten beeinflusst, die selbst wiederum durch die soziale Umgebung mitbestimmt sind, aber auch auf diese einwirken. Fortschritte im Schreibprozess und Evaluationen dieser Fortschritte haben selbst wiederum Einflüsse auf die motivationalen und affektiven Zustände des Schreibers.

Abbildung 2.1.3.1: ***Schreibprozessmodell von Hayes (1996), modifizierte Version des Modells von Hayes und Flower (1980).***



In diesem Modell wird bereits deutlich, dass *Schreibkompetenz* ein komplexes Konstrukt ist und Schreiber sich aus unterschiedlichen Gründen als bessere oder schlechtere Schreiber erweisen können, so etwa aufgrund differenter inhaltlicher Sachkenntnisse, unterschiedlichen Sprachwissens, anderer Rezipientenantizipationen oder auch verschiedener motivationaler Lagen.

2.1.4. Neuorientierung im Bildungswesen: Kompetenzwende, empirische Wende und die Bildungsstandards

In den letzten 20 Jahren kam es zur oftmals sogenannten *Kompetenzwende* im Bildungswesen, einem Paradigmenwechsel in der Lehre. Während Unterricht zuvor vorwiegend input-orientiert gestaltet war, rückte der Fokus nun mehr auf eine lernerzentrierte und outcome-orientierte Unterrichtsgestaltung (Böhme, Richter, Stanat, Pant & Köller, 2012; Elke, 2007; Fuchs, 2009; Klieme & Rakoczy, 2008; Scheerens & Bosker, 1997). Dies bedeutete im Wesentlichen ein Abrücken von primär an Sachverhaltsvermittlung orientiertem Unterricht und einer damit verbundenen Überprüfung deklarativen Wissens hin zu einer Unterrichts-gestaltung, die verstärkt der Ausbildung von Kompetenzen dienlich sein sollte.

In der deutschen Bildungspolitik manifestierte sich diese Wende in Form des Konstanzer Beschlusses im Oktober 1997, in welchem die Entwicklung eines Evaluationssystems zur Bestimmung der Qualität der schulischen Bildung beschlossen wurde sowie die „Durchführung regelmäßiger länderübergreifender Vergleichsuntersuchungen zum Lern- und Leistungsstand von Schülerinnen und Schülern ausgewählter Jahrgangsstufen an allgemeinbildenden Schulen“ (KMK, 1997, S. 1). Angestoßen wurde der Konstanzer Beschluss (auch) durch das mittelmäßige Abschneiden von Schülerinnen und Schüler in Deutschland in TIMSS 1995 (Baumert, Bos & Lehmann, 2000). Verstärkt wurde dieses Bild der Mittelmäßigkeit in Folge durch die PISA-Studie im Jahre 2000 (Kunter et al., 2001), deren Ergebnisse eine breite Rezeption und Diskussion in der Öffentlichkeit erfahren haben, weshalb dieses Ereignis auch häufig als *PISA-Schock* bezeichnet wird (Seitz, 2003). Merkens (2006) fasst dies wie folgt zusammen: „Das klassische Vertrauen darauf, dass das, was gelehrt wird, auch gelernt wird, hat sich als nicht haltbar erwiesen“ (S. 17). Aus diesem Grunde wurden Bildungsstandards, welche zu erwartende Kompetenzen und Teilkompetenzen für einen bestimmten Bereich (Fach, Domäne/Fähigkeitskomplex) normativ festlegen, entwickelt und 2003/2004 als länderübergreifend geltende verbindliche Standards verabschiedet (KMK, 2005b).

Die Bildungsstandards wurden von Vertreterinnen und Vertretern der Schulpraxis, der Fachdidaktik und der Bildungsadministration entwickelt. Darüber hinaus diente eine Expertise im Auftrag des Bundesministeriums für Bildung und Forschung aus dem Jahr 2002 als Grundlage, an deren Erarbeitung Expertinnen und Experten der Allgemeinen Erziehungswissenschaft, der Empirischen Bildungsforschung, der Lehr-Lern-Forschung, des Bildungsrechts, der Historisch-Systematischen Erziehungswissenschaft, der Pädagogisch-

Psychologischen Methodenlehre sowie diverser Fachdidaktiken beteiligt waren (Klieme et al., 2007, S. 15). Im Rahmen von Fachtagungen wurde darüber hinaus die Öffentlichkeit in Form von Vertreterinnen und Vertretern aus Wissenschaft, Wirtschaft, Fachdidaktik, Lehrer-, Schüler- und Elternschaft miteinbezogen. Die Standards wurden im Rahmen dieser Tagungen vor der Beschlussfassung diskutiert und ggf. entsprechend modifiziert (KMK 2005b).

Den Bildungsstandards kommen hierbei zwei zentrale Merkmale zu, welche den Grundstein für die Durchführbarkeit von Schulleistungsstudien zur Überprüfung dieser Standards legten: Kompetenzorientierung und länderübergreifende Verbindlichkeit (Klieme et al., 2007).

Die Definition von Ziel-*Kompetenzen* im Rahmen der Bildungsstandards spiegelt hierbei den Perspektivenwechsel weg von der Input-Orientierung hin zur Outcome-Orientierung und Lernerzentrierung wider. Diese Outcome-Orientierung dient dabei als Grundlage für die Entwicklung und den Einsatz psychologischer Messverfahren (Klieme et al., 2007; Köller, 2008; 2010) und steht im Einklang mit den modernen Ansätzen der Fachdidaktik, in welchen der Fokus auf dem *Können* der Schülerinnen und Schüler liegt (vgl. Kapitel 2.1.2. und 2.1.3).

Aufgrund der länderübergreifenden Verbindlichkeit wurde ein Bezugsrahmen geschaffen, der unabhängig von länderspezifischen Curricula Gültigkeit besitzt und somit eine systematische Vergleichbarkeit von Schülerleistungen über Ländergrenzen hinaus gewährleistet.

In Tabelle 2.1.4.1 sind die für das Ende der Sekundarstufe I für den Kompetenzbereich *Schreiben* im Fach *Deutsch* verabschiedeten Bildungsstandards gemäß den Beschlüssen der Kultusministerkonferenz vom 4.12.2003 und 15.10.2004 (KMK, 2004, 2005a) gelistet.

Tabelle 2.1.4.1: ***Bildungsstandards für den Kompetenzbereich Schreiben für den Hauptschulabschluss und den Mittleren Schulabschluss.***

	Hauptschulabschluss	Mittlerer Schulabschluss
über Schreib-fertigkeiten verfügen	• Texte in gut lesbarer handschriftlicher Form und in einem der Situation entsprechenden Tempo schreiben • Texte dem Zweck entsprechend und adressatengerecht gestalten, sinnvoll aufbauen und strukturieren: z. B. Blattaufteilung, Rand, Absätze • Textverarbeitungsprogramme und ihre Möglichkeiten nutzen: z. B. Formatierung, Präsentation • Formulare ausfüllen	• Texte in gut lesbarer handschriftlicher Form und in einem der Situation entsprechenden Tempo schreiben • Texte dem Zweck entsprechend und adressatengerecht gestalten, sinnvoll aufbauen und strukturieren: z. B. Blattaufteilung, Rand, Absätze • Textverarbeitungsprogramme und ihre Möglichkeiten nutzen: z. B. Formatierung, Präsentation • Formulare ausfüllen
richtig schreiben	• Grundregeln der Rechtschreibung und Zeichensetzung kennen und anwenden • häufig vorkommende Wörter – auch wichtige Fachbegriffe und Fremdwörter – richtig schreiben • individuelle Fehlerschwerpunkte erkennen und Fehler durch Anwendung von Rechtschreibstrategien vermeiden: z. B. Ableiten, Wortverwandtschaften suchen, grammatisches Wissen nutzen	• Grundregeln der Rechtschreibung und Zeichensetzung sicher beherrschen und häufig vorkommende Wörter, Fachbegriffe und Fremdwörter richtig schreiben • individuelle Fehlerschwerpunkte erkennen und mit Hilfe von Rechtschreibstrategien abbauen, insbesondere Nachschlagen, Ableiten, Wortverwandtschaften suchen, grammatisches Wissen anwenden
Texte planen und entwerfen	• den Schreibauftrag verstehen • einen Schreibplan entwickeln • Informationsquellen nutzen: z. B. Bibliotheken, Nachschlagewerke, Zeitungen, Internet • Stoffsammlung erstellen, Informationen ordnen: z. B. Mindmap	• gemäß den Aufgaben und der Zeitvorgabe einen Schreibplan erstellen, sich für die angemessene Textsorte entscheiden und Texte ziel-, adressaten- und situationsbezogen, ggf. materialorientiert konzipieren • Informationsquellen gezielt nutzen, insbesondere Bibliotheken, Nachschlagewerke, Zeitungen, Internet • Stoffsammlung erstellen, ordnen und eine Gliederung anfertigen: z. B. numerische Gliederung, Cluster, Ideenstern, Mindmap, Flussdiagramm

Texte schreiben	• gedanklich geordnet schreiben • formalisierte lineare Texte/ nichtlineare Texte verfassen: z. B. sachlicher Brief, Lebenslauf, Bewerbungsschreiben, Ausfüllen von Formularen, Schaubild, Diagramm, Tabelle • grundlegende Schreibfunktionen umsetzen: erzählen, berichten, informieren, beschreiben, appellieren, argumentieren • produktive Schreibformen nutzen: z. B. umschreiben, weiterschreiben, ausgestalten • kreative Schreibformen nutzen: z. B. Figurengeschichten, Verwandlungsgeschichten, Schreiben zu Bildern • Inhalte verkürzt wiedergeben • wesentliche Informationen aus linearen und nichtlinearen Texten zusammenfassen • wesentliche Gestaltungsmittel untersuchen und darstellen • Argumente finden und formulieren • Argumente gewichten und Schlüsse ziehen • begründet Stellung beziehen • Texte sprachlich gestalten: strukturiert, verständlich und zusammenhängend schreiben • Texte mit Hilfe von neuen Medien verfassen: z. B. Textverarbeitungs- und Mailprogramme	• formalisierte lineare Texte/ nichtlineare Texte verfassen: z. B. sachlicher Brief, Lebenslauf, Bewerbung, Bewerbungsschreiben, Protokoll, Annonce/Ausfüllen von Formularen, Diagramm, Schaubild, Statistik • zentrale Schreibformen beherrschen und sachgerecht nutzen: informierende (berichten, beschreiben, schildern), argumentierende (erörtern, kommentieren), appellierende, untersuchende (analysieren, interpretieren), gestaltende (erzählen, kreativ schreiben) • produktive Schreibformen nutzen: z. B. umschreiben, weiterschreiben, ausgestalten • Ergebnisse einer Textuntersuchung darstellen: z. B. – Inhalte auch längerer und komplexerer Texte verkürzt und abstrahierend wiedergeben – Informationen aus linearen und nichtlinearen Texten zusammenfassen und so wiedergeben, dass insgesamt eine kohärente Darstellung entsteht – formale und sprachlich stilistische Gestaltungsmittel und ihre Wirkungsweise an Beispielen darstellen – Textdeutungen begründen – sprachliche Bilder deuten – Thesen formulieren – Argumente zu einer Argumentationskette verknüpfen – Gegenargumente formulieren, überdenken und einbeziehen – Argumente gewichten und Schlüsse ziehen – begründet Stellung nehmen • Texte sprachlich gestalten – strukturiert, verständlich, sprachlich variabel und stilistisch stimmig zur Aussage schreiben – sprachliche Mittel gezielt einsetzen: z. B. Vergleiche, Bilder, Wiederholung • Texte mit Hilfe von neuen Medien verfassen: z. B. E-Mails, Chatroom

Texte überarbeiten	• eigene und fremde Texte hinsichtlich Aufbau, Inhalt und Formulierungen revidieren • Verfahren zur Überprüfung der sprachlichen Richtigkeit kennen und nutzen	• Aufbau, Inhalt und Formulierungen eigener Texte hinsichtlich der Aufgabenstellung überprüfen (Schreibsituation, Schreibanlass) • Strategien zur Überprüfung der sprachlichen Richtigkeit und Rechtschreibung anwenden
Methoden und Arbeitstechniken	• Notizen machen, Stichpunkte sammeln und ordnen • Arbeitsschritte festlegen • Texte formal gestalten/überarbeiten: z. B. Blattaufteilung, Rand, Absätze, Schriftbild • Texte optisch gestalten • unterschiedliche Informationsquellen nutzen • mit Textverarbeitungs- und Mailprogrammen umgehen • Schreibkonferenzen durchführen • Wörterbücher und Nachschlagewerke nutzen • zentrale Arbeitstechniken kennen und selbstständig anwenden: Abschreiben (von Texten), Aufschreiben, Nachschlagen • Portfolio (selbst verfasste und für gut befundene Texte, Kriterienlisten, Stichwortkonzepte, Selbsteinschätzungen, Beobachtungsbögen von anderen, vereinbarte Lernziele etc.) anlegen und nutzen	• Vorgehensweise aus Aufgabenstellung herleiten • Arbeitspläne/Konzepte entwerfen, Arbeitsschritte festlegen: Informationen sammeln, ordnen, ergänzen • Fragen und Arbeitshypothesen formulieren • Texte inhaltlich und sprachlich überarbeiten: z. B. Textpassagen umstellen, Wirksamkeit und Angemessenheit sprachlicher Gestaltungsmittel prüfen • Zitate in den eigenen Text integrieren • Einhaltung orthografischer und grammatischer Normen kontrollieren • mit Textverarbeitungsprogrammen umgehen • Schreibkonferenzen/Schreibwerkstatt durchführen • Portfolio (selbst verfasste und für gut befundene Texte, Kriterienlisten, Stichwortkonzepte, Selbsteinschätzungen, Beobachtungsbögen von anderen, vereinbarte Lernziele etc.) anlegen und nutzen

Da die Überprüfung der Bildungsstandards bzw. des Erfolgs der Umsetzung dieser Standards mittels vergleichender Studien in Form von Schülerkompetenzmessungen und somit empirisch erfolgen sollte (KMK, 1997; Klieme et al., 2007), bezeichnet man diese Veränderung im Bildungswesen auch als *empirische Wende* (Buchhaas-Birkholz, 2009; Helmke, 2003).

Dieser Wandel erfolgte somit auf zweierlei Ebenen, zum einen auf theoretisch-konzeptioneller Ebene im Sinne der Kompetenzwende mit dem Ziel, den Fokus im Unterricht verstärkt auf Kompetenzvermittlung zu legen, zum anderen auf einer praktisch-diagnostischen Ebene im Sinne der empirischen Wende mit dem neuen Anspruch, den Kompetenzerwerb bzw. dessen Erfolg systematisch empirisch zu erfassen und zu überprüfen.

Für das Unterrichtsfach *Deutsch* erfolgte dieser fächerübergreifende Paradigmenwechsel in einer Zeit, in der ohnehin zumindest für den Kompetenzbereich *Schreiben* die aktuellen fachdidaktischen Strömungen die Prozesse stärker fokussierten und Schreibkompetenzen in das Zentrum der Aufmerksamkeit traten (vgl. Kapitel 2.1.2.).

2.2. Begriffliche Klärung: *Schreibkompetenz*

2.2.1. Zum Kompetenzbegriff

Der Begriff *Kompetenz* wird häufig so aufgefasst und interpretiert, als handle es sich hierbei um die maximale Leistungsfähigkeit von Personen bzw. Personengruppen (zu einem bestimmten Entwicklungszeitpunkt) (Wise, 2009; Wolf & Smith, 1995). Diese Interpretation steht weitgehend im Einklang mit klassischen Kompetenzbegriffen in der Sprachwissenschaft und der Psychologie.

So führte in der Linguistik Noam Chomsky (1962, 1965, 2000) den Kompetenzbegriff im Rahmen einer Unterscheidung zwischen sprachlicher *Kompetenz* und sprachlicher *Performanz* ein. Die Kompetenz beruht hierbei auf den abstrakten Kenntnissen eines Sprecher-Hörers, welche es diesem ermöglichen, aus einem endlichen Inventar an sprachlichen Einheiten wie Lauten oder Wörtern sowie einer begrenzten Anzahl an Regeln unendlich viele grammatisch korrekte Sätze zu generieren. Der Begriff *Performanz* bezieht sich hingegen auf die konkrete, jeweils aktuelle Verwendung der Sprache.

Auch R. W. White (1959), der den Kompetenzbegriff in der Motivationspsychologie einführte, unterscheidet zwischen *Kompetenz* und *Performanz*. Auch ihm zufolge handelt es sich bei Kompetenzen um individuelle Dispositionen, welche dem Handeln, den Performanzen, zugrunde liegen.

Die Unterscheidung zwischen *Kompetenz* und *Performanz*, wie von Chomsky und White vertreten, impliziert jedoch, dass stets nur Performanzen sichtbar sind und Kompetenzen niemals direkt beobachtet und gemessen werden können (Erpenbeck & Rosenstiel, 2007; Shohamy, 1996; Wilkening, 2006). Auch wenn Kompetenzen Performanzen mitbedingen, so sind Performanzen jedoch auch durch andere Aspekte beeinflusst, vor allem auch durch motivationale Aspekte, welche im Rahmen des klassischen Kompetenzbegriffs nicht Teil der Kompetenz sind (Klug, 2007).

Demgegenüber steht ein Kompetenzbegriff, wie etwa von Weinert definiert, gemäß welchem Kompetenzen verstanden werden als „die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können“ (Weinert, 2001, S. 27f.). Ein auf diese Weise bestimmter Kompetenzbegriff bezieht motivationale Aspekte mit ein. Auch zahlreiche andere Autoren stärken einen Kompetenzbegriff, der sowohl *Leistung* als auch *Bereitschaft* umfasst (Becker-Mrotzek & Schindler, 2007; Klieme & Hartig, 2008; Marquard, 1981; Spencer & Spencer, 1993).[4] Schott und Azizi Ghanbari (2008) gehen einen Schritt weiter und schlagen vor, die Kompetenz durch eine Menge von Aufgaben zu beschreiben, durch deren Lösen man das Vorhandensein der Kompetenz nachweist. In diesen Vorschlägen, *Kompetenz* zu definieren, messbar zu machen und letztendlich auch real zu erheben, wird die klassische *Kompetenz-Performanz*-Unterscheidung somit aufgegeben (Erpenbeck & Rosenstiel, 2007).

Eine weitere Unterscheidung, die aufs Engste damit verbunden ist, welchem Kompetenzbegriff man folgt, ist die zwischen *maximalem* und *typischem* Verhalten, welches die Getesteten in Kompetenztests zeigen sollten. Folgt man dem klassischen Kompetenzbegriff, erwartet man, dass die Testteilnehmerinnen und -teilnehmer maximale Leistung zeigen; nur so kann anhand der Performanz auf die Kompetenz rückgeschlossen werden. Hingegen erwartet man unter Einbeziehung motivationaler Aspekte in den Kompetenzbegriff, wie von Weinert (2001) oder Klieme und Hartig (2008) vertreten, lediglich typisches Verhalten der Testpersonen (Asseburg 2011).

In der empirischen Bildungsforschung und somit auch im Rahmen von Studien zur Kompetenzerfassung wie die im Folgenden vorgestellten wird ein Kompetenzbegriff zugrunde gelegt, welcher sich an der Definition von Weinert (2001) orientiert (Artelt et al., 2007; Baumert, Stanat & Demmrich, 2001; Becker-Mrotzek & Böttcher, 2014).

[4] Die Einbeziehung motivationaler Aspekte spiegelt sich auch in dem kognitionspsychologisch orientierten Schreibprozessmodell von Hayes und Flower wider (vgl. Kapitel 2.1.3.).

2.2.2. Der Begriff *Schreibkompetenz*

Bei *Schreibkompetenz* handelt es sich nicht um ein klar umrissenes, wohl definiertes Konzept (Sieber, 2006). Es finden sich in der Fachliteratur eine Vielzahl von Definitionen und Spezifikationen von *Schreibkompetenz* (Becker-Mrotzek, Jost, Knopp & Grabowski, 2011; Bereiter, 1980; M. Fix, 2006; Grabowski, Blabusch und Lorenz, 2007; Merz-Grötsch, 2001; Molitor-Lübbert, 1996; NAEP, 2011b; Ossner 1995; Rehder, 2011; Richter, 2008). Allerdings lassen sich unter diesen prinzipiell zwei verschiedene Basis-Spezifikationen ausmachen:

a) Die Fähigkeit, einen stilistisch angemessenen zusammenhängenden Text zu verfassen, in welchem die relevanten Inhalte vollständig/angemessen/hinreichend/plausibel (re)produziert werden (protypisch: Richter, 2008).
b) Die Fähigkeit, einen stilistisch angemessenen sowie orthografisch und grammatikalisch korrekten zusammenhängenden Text zu verfassen, in welchem die relevanten Inhalte vollständig/angemessen/hinreichend/plausibel (re)produziert werden (prototypisch: NAEP, 2011b).

Eine weitere Differenzierung findet sich dahingehend, dass unter *Fähigkeit* die rein kognitive[5] Fähigkeit verstanden werden kann oder die kognitive und graphomotorische Fähigkeit (Antos, 1996; Merz-Grötsch, 2001; Bereiter, 1980).

Da im Rahmen der Beurteilung und Messung schulischer Kompetenzen am Ende der Sekundarstufe I im allgemeinbildenden Schulsystem graphomotorische Fähigkeiten als gegebene physiologische Voraussetzung angesehen werden können, spielen diese im Rahmen von Schreibassessmentstudien in dieser Altersstufe keine Rolle. Die beiden spezifizierten Definitionstypen unterscheiden sich folglich im Umfang dessen, welche Teilfähigkeiten unter *Schreibkompetenz* subsumiert werden sollen; sind dies stilistische und inhaltliche Fähigkeitsaspekte (wie unter a) oder werden (wie unter b) zusätzlich Fähigkeiten zur Anwendung sprachlicher Richtigkeitsnormen inkludiert.

Welcher Konzeption man hierbei folgt, ist eine Frage des Anspruches an den Begriff und die Messung von Schreibkompetenz. Ist der primäre Anspruch, mit dem Konstrukt *Schreibkompetenz* ein eindimensionales Konstrukt zu erfassen (beispielsweise in Folge von vorgeschalteten Dimensionsanalysen), so erweist sich eine Definition wie unter a), d. h. unter Ausklammerung von Fähigkeiten, welche die sprachliche Richtigkeit betreffen, als geeignet

[5] Der Begriff *kognitiv* im hiesigen Sinne inkludiert auch motivationale Aspekte (vgl. Kapitel 2.1.3. & 2.2.1.).

(vgl. im die folgenden Ausführungen unter 2.3.5.). Allerdings kann – wie auch in den drei letztgenannten Studien – das Konstrukt *Schreibkompetenz* extern festgelegt sein, beispielsweise durch den Anspruch, alle textqualitätsbestimmenden Aspekte zu erfassen. Welche Aspekte dies sind, wird durch den theoretischen und praktischen Bezugsrahmen festgelegt.

Da es sich bei den in dieser Arbeit vorgestellten Studien um Studien zur Überprüfung der Bildungsstandards handelt, orientiert sich der zugrunde zu legende Schreibkompetenzbegriff an den in diesen festgelegten Fähigkeitsaspekten. Dort werden sowohl stilistisch-strukturelle Schreibfähigkeiten (wie beispielsweise „Texte sprachlich gestalten: strukturiert, verständlich, sprachlich variabel und stilistisch stimmig zur Aussage schreiben“ oder „sprachliche Mittel gezielt einsetzen: z. B. Vergleiche, Bilder, Wiederholung“) und inhaltliche Schreibfähigkeiten (wie etwa „Argumente gewichten und Schlüsse ziehen“ oder „Inhalte auch längerer und komplexerer Texte verkürzt und abstrahierend wiedergeben“) als auch Rechtschreibfähigkeiten (u. a. „Grundregeln der Rechtschreibung und Zeichensetzung sicher beherrschen und häufig vorkommende Wörter, Fachbegriffe und Fremdwörter richtig schreiben“) angeführt (KMK, 2005b, S. 11–12; vgl. Tabelle 2.1.4.1).

Darüber hinaus orientiert sich ein wesentlicher Teil des Auswertungsschemas (*holistisches Kodierschema*, vgl. Kapitel 3.4.) an den Auswertungsmethoden des National Assessment of Educational Progress (NAEP). Hier wird Schreibkompetenz definiert als „die Fähigkeit, Ideen zu entwickeln und auszuarbeiten, die eigenen Gedanken zu organisieren und in grammatikalisch korrekter Prosa zu schreiben“ (NAEP, 2011b, S. vi; Übersetzung T.C.).

Die für diese Arbeit angemessene Definition von *Schreibkompetenz* ist somit (entsprechend obiger Version b), jedoch unter expliziter Ausklammerung graphomotorischer Fähigkeiten):

> Die (kognitive) Fähigkeit, einen stilistisch angemessenen sowie orthografisch und grammatikalisch korrekten zusammenhängenden Text zu verfassen, in welchem die relevanten Inhalte vollständig/angemessen/hinreichend/plausibel (re)produziert werden.

2.3. Empirische Schreibleistungsforschung

2.3.1. Frühe Aufsatzstudien und die Subjektivität von Schreibleistungs-beurteilungen

Untersuchungen in den 1960er und 1970er Jahren konnten zeigen, dass die Bewertung von Aufsätzen durch verschiedene Bewertungspersonen (in der Regel Lehrerinnen und Lehrer) nicht einheitlich erfolgt. Schröter (1971) ließ 617 Schüleraufsätze von jeweils 10 oder mehr Lehrpersonen mit Schulnoten bewerten. Dabei zeigte sich bei fast allen Texten (95 %) eine Spannweite von mindestens 3 Schulnoten in den Bewertungen, in über 10 % der Fälle sogar ein Bewertungsspektrum über 5 Schulnoten. Weiss (1965; 1966) lieferte Evidenz, dass Aufsatzbewertungen sich mit gegebenen Hintergrundinformationen zu den Schreibendem positiv bzw. negativ beeinflussen lassen. Weitere Studien wiesen beeinflussende Effekte der Handschrift, der Textlänge, der Rechtschreibung bei rein inhaltlicher Textbeurteilung oder auch textexterner Faktoren wie der Schülerbeliebtheit nach (Briggs, 1970; Ingenkamp, 1971, Marshal, 1967; Weber, 1973).

In Folge dieser Befunde wurden Bewertungskriterien gefordert und entwickelt, welche die Urteilerübereinstimmung bei der Bewertung von Aufsätzen enorm erhöhten (u. a. Beck, 1974; Grzesik & Fischer, 1984; Nussbaumer 1991). Dennoch blieben Text- und Schreibleistungs-beurteilungen weiterhin (verglichen mit Bewertungen in anderen Bereichen) mit einem überdurchschnittlichen Maß an Subjektivität bzw. Nichtübereinstimmung verbunden. Fix und Melenk (2002) weisen in diesem Zusammenhang auf die Unvermeidbarkeit einer gewissen Restsubjektivität hin:

> „Unter textrezeptionstheoretischen Gesichtspunkten können Textbewertungen nicht frei von subjektiven Einflüssen des Bewerters sein, wenn man davon ausgeht, dass ein Text vom Leser selektiv wahrgenommen und individuell rezipiert wird.“ (S. 47)

Durch den Einsatz solcher Beurteilungssysteme sowie ergänzender Beispieltexte konnte jedoch „die Reliabilität (…) so sehr gesteigert werden (…), dass sie nahe an die Größenordnung heran kommt, die für formelle Tests gefordert wird“ (Birkel, 2003, S. 47).

Die unterschiedlichen Entwicklungen im Bereich der Bewertungskriterien lassen sich hierbei prinzipiell in zwei verschiedene Beurteilungstypen oder Beurteilungsstrategien unterscheiden. Diese werden im folgenden Subkapitel erläutert sowie ihre Vor- und Nachteile gegenübergestellt.

2.3.2. Beurteilungsverfahren, ihre Vor- und Nachteile

Im Rahmen der empirischen Schreibforschung ist ein zentrales Problem, wie Schreibleistungen geeignet beurteilt werden sollen. Hier stehen sich prinzipiell zwei Beurteilungsstrategien gegenüber, eine holistische und eine analytische. Bei holistischen Beurteilungen handelt es sich um ganzheitliche Textbewertungen, im Rahmen derer alle Textqualitätsaspekte berücksichtigt werden und in ein Urteil miteinbezogen werden. Bei analytischen Beurteilungen handelt es sich um Bewertungen einzelner Textqualitätsaspekte im Rahmen getrennter Einzelurteile (bspw. *richtige Textsorte*, *einheitliche Perspektive* etc.); bei analytischen Beurteilungssystemen liegt somit immer ein Katalog an zu bewertenden Einzelkriterien vor, insofern beabsichtigt wird, alle (relevanten / zentralen) Textqualitätsmerkmale zu erfassen. Beide Beurteilungsverfahren weisen diverse Vor- und Nachteile auf.

Geht man auf konzeptueller Ebene davon aus, dass es sich bei *Schreibkompetenz* um ein eindimensionales Konstrukt handelt, wird eine ganzheitliche Erfassung der Textqualität diesem Verständnis besser gerecht als eine kriteriale Erfassung; das holistische System weist unter dieser Hintergrundannahme somit ein höheres Maß an (Inhalts- und Konstrukt-) Validität auf.[6] Eine holistische Erfassung vermeidet hierbei eine mögliche lückenhafte Erfassung, welche ein analytisches System potentiell mit sich bringt; analytische Systeme erfassen die Textqualität in Form einer begrenzten Anzahl an Kriterien; weist ein Text Merkmale auf, die nicht durch diese Kriterien abgedeckt sind, werden diese im Rahmen der Beurteilung nicht berücksichtigt. Im Rahmen einer holistischen Beurteilung kann sich auf den Text als Ganzes bezogen und somit kontextuellen Besonderheiten Rechnung getragen werden, welche aus dem jeweiligen Text hervorgehen (E. M. White, 1984, 1985; Weigle, 2002).

Eine analytische Betrachtung ermöglicht hingegen eine Erfassung von Details. Dies ist etwa bei Leistungsrückmeldungen im Rahmen des Unterrichts sinnvoll, wenn konkrete unterrichtsgestaltende oder Fördermaßnahmen angeschlossen werden. Hier werden differenzierte und präzise Informationen bereitgestellt, hinsichtlich welcher Teilaspekte (bspw. Wortschatz, Grammatik, Textsortenkenntnis) individuelle Schwächen und Stärken bei einem Schreibenden vorhanden sind. Auch kann somit potentiell asymmetrischen Schreibprofilen Rechnung getragen werden und diese können in ihrer Schiefe erfasst werden (Beck, 1979; Hamp-Lyons, 1991; Hofen, 1980; A. Neumann, 2007; Weigle, 2002). Darüber hinaus ist die Erfassung separater Kriterien eine notwendige Voraussetzung, um

[6] Für nähere Erläuterungen und Ausführungen zur Validität und zu Validitätsaspekten vgl. Kapitel 5.

Dimensionsanalysen hinsichtlich des Konstrukts *Schreibkompetenz* durchzuführen und somit auch zu prüfen, ob die Annahme eines eindimensionalen Konstrukts, wie es Befürworter eines holistischen Systems oftmals voraussetzen, einer empirischen Prüfung standhält.

Hinsichtlich der Modellierung des Konstrukts *Schreibkompetenz* sowie bei der Ermittlung eines singulären Schreibkompetenzwertes generell erweisen sich holistische Beurteilungssysteme als besser geeignet. Würde man analytische Kriterien heranziehen, müssten diese zunächst verrechnet werden. Dies ist jedoch mit einigen theoretischen und praktischen Problemen und Herausforderungen verbunden. So ist unklar, wie diese Verrechnung aussehen sollte. Einfache Summenscores scheinen dem Konstrukt nicht gerecht zu werden; so gibt es Kriterien, die konsensuell wichtiger und zentraler für ein gelungenes Schreibprodukt sind (bspw. *richtige Textsorte*) als andere (bspw. *formale Gliederung in Absätze*). Daher müsste eine explizite Gewichtung dieser Aspekte erfolgen. Diese Gewichtung müsste jedoch entweder normativ gesetzt oder empirisch ermittelt werden. Letzteres ist jedoch wiederum nur möglich, wenn ein Abgleich mit einem bereits bestehenden Urteil zur Schreibkompetenz insgesamt vorliegt, wie es etwa im Rahmen einer holistischen Beurteilung hätte generiert werden können. Für eine normative Setzung hingegen stellen sich Anschlussfragen wie, ob die Gewichtung aufgabenübergreifend oder aufgabenspezifisch erfolgen sollte, sowie das generelle Problem, wie die einzelnen Kriterien zu gewichten sind, um das Konstrukt hinreichend valide zu erfassen.

Hinsichtlich einer möglichst zuverlässigen Beurteilung wurde für einige Studien nachgewiesen, dass eine analytische Beurteilung reliabler ist, vor allem für den Fall, dass ungeschulte Raterinnen und Rater die Beurteilung durchführen, und sich ähnlich hohe oder höhere Urteilerübereinstimmungen in holistischen Verfahren nur nach intensiver Raterschulung erreichen lassen (Bauer, 1981; Grzesik & Fischer, 1984; Swartz et al., 1999; Weigle, 2002; Weir, 1990).

Ein Faktor, der für die höhere Reliabilität analytischer Kriterien verantwortlich ist, liegt in der konkreten Operationalisierung begründet. Analytische Kriterien werden in den meisten Fällen dichotom erfasst. Zwar ist es prinzipiell auch möglich, ordinale Kriterien im Rahmen einer analytischen Kodierung zu verwenden, aus zeitökonomischen und den im Folgenden erläuterten psychometrischen Gründen wird jedoch meist darauf verzichtet. Holistischen Beurteilungssystemen liegen stets ordinale Skalen zugrunde. So werden feingliedrige Abstufungen bei der Leistungsbeurteilung ermöglicht, obwohl jeweils nur ein Urteil (pro Text) vorliegt. Ordinale Kriterien sind allerdings stärker Beurteilertendenzen wie etwa der

Tendenz zur Mitte ausgesetzt als dichotome (Jonkisz, Mossbrugger & Brandt, 2012; Tiemann & Körbs, 2014).

Darüber hinaus gibt es einige praktische Aspekte, hinsichtlich derer sich die holistischen und analytischen Beurteilungssysteme unterscheiden. So ist einerseits die Textbeurteilung nach einem analytischen System in der Regel zeitaufwändiger (A. Neumann, 2007; Weigle, 2002). Andererseits weist das analytische System gegenüber dem holistischen eine deutlich kürzere ‚Einarbeitungszeit' auf. Des Weiteren operieren holistische Systeme häufig mit Oberbegriffen, Zusammenfassungen und Abstraktionen, deren angemessene Interpretation nicht ohne Zusatzinformationen zu gewährleisten ist (Weigle, 2002); aus diesem Grund werden holistische Systeme in der Regel mit prototypischen Beispieltexten (*Benchmarks*) ergänzt.

Zahlreiche Autoren schlagen aufgrund der diversen Vor- und Nachteile der Systeme einen kombinierten Einsatz von analytischer und holistischer Textbeurteilung vor (Bachman & Palmer, 1996; M. Fix & Melenk, 2002; Harsch et al., 2007; W. Hartmann & Lehmann, 1987; Lehmann, 1990, 1994; A. Neumann, 2012; Nussbaumer 1991).

2.3.3. Kompetenzskalen und Kompetenzstufenmodelle

Um ausgehend von den Beurteilungen von Texten auf die Schreibkompetenz zu schließen, ist es notwendig, die Leistungen auf einer Kompetenzskala abzubilden. Solche Skalierungen werden durch die Anwendung von IRT-Modellen erreicht. IRT-Modelle ermöglichen es, anhand von manifesten Leistungsdaten auf eine latente Variable (wie etwa *Schreibkompetenz*) zurück zu schließen. Dabei werden zunächst bestimmte Annahmen über die mit den einzelnen Aufgaben (Items) erfassten Teilleistungen, sowie über die beteiligten Personen spezifiziert. So muss etwa vorab definiert sein, ob alle Items denselben Beitrag zur modellierten Variablen betragen und ein einparametrisches Modell wie beispielsweise das Rasch-Modell (Rasch, 1960) angewandt werden soll oder ob die Items unterschiedlich stark diskriminieren sollen, wofür ein zweiparametrisches Modell (u. a. Birnbaum, 1968) gewählt werden müsste. Darüber hinaus können Personeneigenschaften, von denen angenommen wird, dass sie leistungsbeeinflussend sein könnten (bei Schülerinnen und Schülern etwa der Besuch einer Schule, die einer bestimmten Schulform zuzuordnen ist), im Hintergrundmodell spezifiziert werden, um im Rahmen der Schätzverfahren berücksichtigt werden zu können (Hartig, Jude & Wagner, 2008; Hartig & Kühnbach, 2006; Hornke et al., 2011). Unter gegebenen Spezifikationen werden im Rahmen des Modells Item- und Personenparameter wie die

Schwierigkeit der einzelnen Teilaufgaben oder die Personenfähigkeit geschätzt. Modellgütemaße spezifizieren, wie gut das theoretische Modell und die empirischen Daten zusammenpassen.

Als Ergebnis des Skalierungsverfahrens wird jeder Person ausgehend von den manifesten Leistungsdaten ein Kennwert auf einer kontinuierlichen metrischen Skala zugeordnet, der sich als Personenfähigkeit hinsichtlich des modellierten latenten Konstrukts (z. B. *Schreibkompetenz*) interpretieren lässt. Oftmals ist es jedoch sinnvoll, diese kontinuierliche Skala in verschiedene Kategorien einzuteilen, um bestimmte Kompetenzstände besser deuten und einordnen zu können. Diese Einteilung leisten Kompetenzstufenmodelle.

Kompetenzstufenmodelle unterteilen eine gegebene metrische Kompetenzskala in verschiedene Kategorien, die Stufen des Modells. Bei diesen Stufen handelt es sich um inhaltlich distinkte Kategorien, die sich als Niveaustufen interpretieren lassen (u. a. Hartig, 2004; Hofmeister, 2005; PISA-Konsortium, 2004). Kompetenzstufenmodelle im Rahmen der Überprüfung des Erreichens der Bildungsstandards orientieren sich hierbei an diesen und definieren die Stufen im Hinblick auf das Erreichen/Nichterreichen bzw. unterschiedliche Grade des Erreichens der Standards (Bremerich-Vos et al., 2010; Bremerich-Vos, Böhme, Krelle, Weirich & Köller, 2012; Pant, Böhme & Köller, 2012).[7]

Bildungsstandardorientierte Kompetenzstufenmodelle im Bereich *Deutsch* liegen bisher für *Lesen*, *Zuhören*, *Rechtschreiben* und *Sprache und Sprachgebrauch untersuchen* sowohl für den Primarbereich als auch für die Sekundarstufe I vor (IQB, 2015). Die Genese des Kompetenzstufenmodells bzw. der Kompetenzstufenmodelle *Schreiben* für die Sekundarstufe I ist Gegenstand dieser Arbeit und wird am Ende des Kapitels 3 beschrieben und erläutert.

2.3.4. Schreibleistungsstudien im Large-Scale-Bereich

Die empirische Wende (vgl. Kapitel 2.1.4.) legte den Grundstein für großangelegte, systematisch nach wissenschaftlichen Kriterien durchgeführte Untersuchungen zur Kompetenzerfassung im Large-Scale-Bereich.

Eine der ersten großangelegten nationalen Bildungsstudien im Large-Scale-Bereich war die DESI-Studie (DESI = Deutsch Englisch Schülerleistungen International), welche von der

[7] Vgl. für detaillierte Ausführungen am konkreten Beispiel der Entwicklung der Kompetenzstufenmodelle *Schreiben* die Kapitel 3.7 und 3.8.

KMK als Ergänzung zur PISA-Studie in Auftrag gegeben und im Schuljahr 2003/2004 durchgeführt wurde. Hierbei wurden rund 11.000 Schülerinnen und Schüler der neunten Jahrgangsstufe zu zwei Messzeitpunkten (Schuljahresbeginn und -ende) getestet. Für den Bereich *Deutsch Schreiben* kamen vier verschiedene Schreibaufgaben, allesamt der Textsorte *Brief* zuzuordnen, zum Einsatz. Zur Beurteilung der Texte wurden dabei mehrere teilweise dichotome, teilweise mehrstufige analytische Kriterien (u. a. *Textaufbau*, *Rechtschreibung*, *Grammatik*, *Wortwahl*) sowie eine fünfstufige holistische Skala *Gesamteindruck* verwendet. (Klieme 2006; Klieme et al. 2006; A. Neumann 2007, 2014).

Bereits vor der empirischen Wende in Deutschland wurden 1985 im Rahmen der Hamburger Aufsatzstudie, welche an die internationale Aufsatzstudie IEA angeschlossen war, Texte zu vier Aufgaben von 1340 Schülerinnen und Schülern erhoben und systematisch bewertet; zur Beurteilung kamen hier erstmals sowohl analytische Kriterien (*Stil*, *Organisation*, *formale Korrektheit*, *Inhalt*) als auch eine holistische Skala zum Gesamteindruck parallel zum Einsatz (W. Hartmann & Jonas, 1996; Lehmann, 1994; Lehmann & Hartmann, 1987).

Darüber hinaus wurden auf Ebene einzelner Bundesländer in den letzten beiden Jahrzehnten in einigen Studien Schreibkompetenzen erfasst, so etwa in Hamburg im Rahmen der LAU-Studie (LAU = Aspekte der Lernausgangslage und der Lernentwicklung), welche 1995 startete und sich in Form einer längsschnittlichen Untersuchung über mehrere Jahre erstreckte. Schreibaufgaben wurden hierbei in den Klassenstufen 5 (1996), 9 (2000) und 11 (2002) eingesetzt (Lehmann & Peek, 1997; Lehmann, Peek, Gänsfuß & Husfeldt, 2002; Lehmann, Hunger, Ivanov, Gänsfuß & Hoffmann, 2004; A. Neumann, 2007, 2014).

Über den deutschsprachigen Raum hinausblickend, sind vor allem die Studien zur Schreibkompetenzerfassung in den USA zu nennen, welche im Rahmen des *National Assessment of Educational Progress* (NAEP) ermittelt werden. NAEP erhebt seit 1998 in regelmäßigen Abständen von vier bis fünf Jahren (1998, 2002, 2007 und 2011) Schreibleistungsdaten von repräsentativen Stichproben[8] von Schülerinnen und Schülern der vierten (nur 1998 und 2002), achten und zwölften Jahrgangsstufe und wertet diese systematisch aus. Diese Auswertung erfolgt bei NAEP textmusterspezifisch, unterschieden wird hierbei zwischen *narrative writing*, *informative writing* und *persuasive writing*. Die Beurteilung erfolgt ausschließlich anhand einer textmusterspezifischen (holistischen) Globalskala (NAEP 1999, 2003, 2008, 2012).

[8] Im Durchgang 2011 waren dies rund 25.000 Schülerinnen und Schüler pro getestete Jahrgangsstufe, in den zurückliegenden Erhebungen waren die Stichproben größer.

In einigen Durchführungs- und Auswertungsaspekten an die Studien von NAEP angelehnt (vgl. Kapitel 3) sind die Studien zur Normierung von Aufgaben zur empirischen Überprüfung des Erreichens der Bildungsstandards im Kompetenzbereich *Deutsch Schreiben*, welche im Auftrag der KMK durch das Institut zur Qualitätsentwicklung im Bildungswesen (IQB) durchgeführt wurden. Im Schuljahr 2006/2007 wurden für den Primarbereich 16 Aufgaben unter Einbeziehung informierender, argumentierender und narrativer Texte pilotiert und 8 Aufgaben schließlich unter Teilnahme von 1806 Schülerinnen und Schülern der dritten und vierten Jahrgangsstufe normiert, dabei kamen sowohl ein analytischer Kriterienkatalog als auch ein holistisches Kodiersystem, bestehend aus einer holistischen Globalskala und drei semiholistischen Subskalen zur ganzheitlichen Erfassung der Aspekte *Inhalt*, *Stil* und *sprachliche Richtigkeit*, zum Einsatz (Böhme, 2012; Winkelmann & Böhme, 2009).

Die Normierungsstudie im Kompetenzbereich *Deutsch Schreiben* für die Sekundarstufe I bildet primäre Grundlage der vorliegenden Arbeit; Rahmen und Durchführung der Studie werden detailliert in Kapitel 3 dargestellt.

2.3.5. Die Struktur von Schreibkompetenzen

Ein Aspekt, der in mehreren der in Kapitel 2.3.4. angeführten Studien untersucht worden ist, ist die innere Struktur von Schreibkompetenz, sprich die Frage danach, ob es sich bei Schreibkompetenz um ein eindimensionales Konstrukt handelt, oder ob sich mehrere in weiten Teilen empirisch voneinander unabhängige Dimensionen ausmachen lassen.

So untersuchte Astrid Neumann (2007) anhand der Daten der DESI-Studie sowie der LAU11/ ULME1-Studie[9] anhand von Strukturgleichungsmodellen auf der Basis der analytischen Schreibleistungsdaten die interne Struktur von *Schreibkompetenz*. Dabei zeigte sich in aufgabenübergreifender Betrachtung eine zweidimensionale Struktur mit den Dimensionen *Semantik/Pragmatik* (basierend auf den Kriterien *Textaufbau*, *Wortwahl*, *Stil* sowie inhaltlichen Merkmalen) und *Sprachsystem* (basierend auf *Rechtschreibung*, *Grammatik* und *Satzkonstruktion*).

[9] Die Zahl 11 bezieht sich hierbei auf die 11. Jahrgangsstufe des allgemeinbildenden Schulsystems. ULME steht für *Untersuchung der Leistung, Motivation und Einstellungen zu Beginn der beruflichen Ausbildung*, im Rahmen der Studie werden diejenigen Schülerinnen und Schüler getestet, die zuvor (in zurückliegenden Erhebungswellen) an der längsschnittlichen LAU-Studie teilgenommen haben, bis zu diesem Testzeitpunkt das allgemeinbildende Schulsystem verlassen haben und inzwischen berufliche Schulen besuchen.

Auch Lehmann und Hartmann (1987) konnten anhand der Daten der Hamburger Aufsatzstudie eine zweidimensionale Struktur nachweisen. Die Autoren unterscheiden hierbei zwischen *Inhalt* und *Aufbau* einerseits und *Mechanics* andererseits. Der *Mechanics*-Faktor bezieht sich hierbei auf allgemeine sprachlich-strukturelle Leistungsaspekte, die sich aufgabenübergreifend als stabil erwiesen.

Böhme, Bremerich-Vos und Robitzsch (2009) erbrachten anhand der Daten der Normierungsstudie im Primarbereich bei einer Kategorisierung der Schreibfähigkeitsaspekte in *Inhalt*, *Stil* und *sprachliche Richtigkeit* anhand der semiholistischen Subskalen (vgl. Kapitel 2.3.4.) ebenfalls Evidenz für eine zweidimensionale Struktur mit *Inhalt* und *Stil* als eine und *sprachliche Richtigkeit* als davon separate zweite Dimension.

2.3.6. Weiteres Vorgehen und Forschungsfragen dieser Arbeit

Wie bereits in Kapitel 1.3. erläutert, besteht der Kern dieser Arbeit aus zwei Teilen, im ersten Teil, den die folgenden beiden Hauptkapitel bilden, wird die Normierungsstudie im Kompetenzbereich *Schreiben* im Fach *Deutsch* am Ende der Sekundarstufe I sowie die anschließende Entwicklung des Kompetenzstufenmodells detailliert beschrieben. In der Studie kamen sowohl holistische als auch analytische Beurteilungsverfahren zum Einsatz, im Rahmen des holistischen Beurteilungssystems auch semiholistische Skalen, wie sie erstmals in vergleichbarer Form in der Normierungsstudie des Primarbereichs (vgl. Kapitel 2.3.4.) verwendet wurden. Ein besonderes Augenmerk wird deshalb auch auf die Urteilerübereinstimmungen der einzelnen Beurteilungssysteme gelegt.

Für die Genese des Kompetenzstufenmodells wird detailliert auf den Prozess des Standard-Settings, ein Verfahren zur Ermittlung der qualitativ unterscheidbaren Kompetenzstufen bzw. deren Grenzen, eingegangen, welches in dieser Form erstmals im deutschsprachigen Raum angewandt wurde.

In Kapitel 4 werden einige zentrale Ergebnisse der Normierungsstudie wie die Kompetenzstufenverteilung der Schülerinnen und Schüler sowie Unterschiede in den Kompetenzständen zwischen Personengruppen berichtet. Diese Ergebnisse werden kapitelabschließend im Zusammenhang mit Befunden anderer sprachlicher Leistungsstudien diskutiert.

Im zweiten Teil dieser Arbeit schließen die drei Forschungsteilstudien an, die zunächst durch ein theoretisches Kapitel zum Konzept der Validität eingeleitet und zu diesem in Bezug gesetzt werden. In der ersten Teilstudie (Kapitel 6) wird der Fragestellung nachgegangen, ob es sich bei *Schreibkompetenz* um ein textmusterunabhängiges Konstrukt handelt oder ob es sich bei *argumentierender*, *informierender* und *narrativer Schreibkompetenz* um differente Konstrukte bzw. Konstruktdimensionen handelt. Dimensionsanalysen im Bereich *Schreiben* beschränkten sich bisher auf die Differenzierung von Dimensionen nach sprachlich-textuellen Aspekten wie *Inhalt*, *Struktur*, *Stil* oder *Rechtschreibung/Grammatik* (vgl. Kapitel 2.3.5.). Auch im Rahmen der Schreibleistungsstudien von NAEP (vgl. Kapitel 2.3.4.), in welchen Schreibkompetenzen textmusterspezifisch erfasst werden, wurden bisher keine Dimensionsanalysen hinsichtlich Textmuster durchgeführt.

Die zweite Teilstudie (Kapitel 7) widmet sich der Frage, ob und inwiefern bei der Messung von Schreibkompetenzen Lesefähigkeiten miterfasst werden. In der Normierungsstudie sowie anderen bisherigen Schreibleistungsstudien (vgl. Kapitel 2.3.4.) gehen den Schreibaufgaben schriftliche Instruktionstexte und ggf. Zusatzinformationen voraus, sodass eine korrekte Bearbeitung der Schreibaufgabe die erfolgreiche Rezeption dieser Vortexte voraussetzt. Sollte das Verständnis der Instruktion aufgrund zu hoher sprachlicher Anforderungen gestört sein, würde sich dies auch im Schreibleistungsergebnis niederschlagen. Konkret wird daher im Rahmen der Teilstudie geprüft, ob der Zusammenhang zwischen in einem (parallel durchgeführten) Lesekompetenztest ermittelten Lesefähigkeiten und den ermittelten Schreibfähigkeiten von der Leseschwierigkeit der Schreibaufgaben abhängen.

Die dritte Teilstudie (Kapitel 8) schließt an die Befunde der Aufsatzstudien der 1960er und 1970er an (vgl. Kapitel 2.3.1.) und geht der Fragestellung nach, ob und inwiefern inhaltliche und stilistische Schreibleistungsbeurteilungen durch orthografische und grammatische Aspekte der zugrunde liegenden Texte beeinflusst sind. Im Vergleich zu den frühen Aufsatzstudien liegen im Rahmen des Schreibassessments veränderte Bedingungen vor, die anonyme Bewertung, der Einsatz von Beurteilungsschemata (vgl. Kapitel 2.3.2) sowie die Schulung in der Anwendung dieser Schemata, sorgen, wie in Kapitel 2.3.1. erläutert, für einen Rückgang der Subjektivität in der Textbeurteilung. Somit wird im Rahmen der Teilstudie untersucht, ob und inwiefern dies auch Einflüsse von beurteilungsirrelevanten Textaspekten reduziert.

3. Normierung im Kompetenzbereich *Schreiben* und Entwicklung der Kompetenzstufenmodelle

In diesem Kapitel wird der Prozess von der Aufgabenentwicklung über die Durchführung der Studie zur Normierung von Aufgaben zur empirischen Überprüfung des Erreichens der Bildungsstandards bis zur Genese der Kompetenzstufenmodelle für den Bereich *Schreiben* dargestellt. Dabei werden die verwendeten Auswertungsinstrumente näher beschrieben und das Verfahren des Standard-Settings detailliert erläutert.

3.1. Aufgabenentwicklung

Die Aufgabenentwicklung für den Kompetenzbereich *Schreiben* wurde in mehreren Etappen durchgeführt und erstreckte sich insgesamt über einen Zeitraum von dreieinhalb Jahren.

Im Zeitraum von März 2007 bis Oktober 2007 wurden von einem Aufgabenentwicklerteam, bestehend aus 14 erfahrenen Lehrkräften der Sekundarstufe I aus mehreren deutschen Bundesländern, im Rahmen der Vorbereitung der Normierung der bildungsstandardbasierten Testaufgaben im Fach *Deutsch* für die Sekundarstufe I Aufgaben für fünf Kompetenzbereiche (*Lesen – Zuhören – Rechtschreiben – Sprache und Sprachgebrauch untersuchen – Schreiben*) entwickelt. Die Begutachtung und Überarbeitung der Aufgaben wurde von einem Expertenteam aus dem Bereich der Deutschdidaktik und der empirischen Bildungsforschung vorgenommen. Die fachdidaktische Betreuung oblag Prof. Dr. Albert Bremerich-Vos und Mitarbeitern, die Expertise der empirischen Bildungsforschung dem IQB. Für den Kompetenzbereich *Schreiben* wurden in diesem Rahmen 26 Aufgaben entwickelt.

Im Zeitraum von Januar 2010 bis Oktober 2010 – nach der Erprobung der ersten Aufgaben – wurden diese überarbeitet; zusätzlich wurden zehn neue Aufgaben unter der fachdidaktischen Betreuung von Prof. Dr. Bremerich-Vos entwickelt.

Im Rahmen der Aufgabenentwicklung wurden sowohl für Textsorten und Textmuster protypische Aufgaben, bspw. die Erzählung eines Erlebnisses, als auch solche Aufgaben generiert, die Elemente aus verschiedenen Textsorten und -mustern kombinieren, bspw. das Schreiben eines Lexikoneintrags, welcher sowohl berichtend-beschreibende als auch narrative Elemente beinhalten sollte. Dieses Vorgehen wurde gewählt, da es zur Zeit der Aufgaben-

entwicklung noch unklar war, ob *Schreibkompetenz* textmusterspezifisch oder textmusterübergreifend modelliert und beschrieben werden wird.[10]

Zwei Beispielaufgaben, eine informierende sowie eine argumentierende, finden sich im Anhang unter A.3.1.1 und A.3.2.2.

Anzumerken ist hierbei, dass die Aufgaben jeweils das Verfassen eines Textes erfordern, der als Grundlage zur Bewertung dient. Die Beurteilung erfolgt somit produktbezogen. Ein Erfassen von einzelnen Schreibprozessen (Planen, Formulieren, Überarbeiten etc.) ist im Rahmen großangelegter standardisierter und zeitlich begrenzter Untersuchungen nicht möglich.[11] Dies steht jedoch nicht im Widerspruch zur Prozessorientierung der fachdidaktischen Schreibforschung der letzten 20 Jahre (vgl. Kapitel 2.1.2. und 2.1.3.). So weist Sieber (2006) auf Folgendes hin:

> „Eine reine prozessbezogene Betrachtungsweise – etwa über Protokolle lauten Denkens – ergibt keinerlei Hinweis darauf, wie sich das Produkt entwickelt und ob das Produkt des so erfassten Prozesses überhaupt der Schreibaufgabe gerecht wird. Dies lässt sich nur aufgrund der produktbezogenen Betrachtung entscheiden. Zudem bleibt unberücksichtigt, dass das Produkt selbst, der bis zu einem Zeitpunkt im Textproduktionsprozess jeweils entstandene Text, nicht nur Ergebnis von Prozessen ist, sondern zugleich diese Prozesse im Folgenden wesentlich beeinflusst, so dass von einer engen Wechselbeziehung zwischen Prozess und Produkt auszugehen ist.“ (S. 211; Originalschreibung hinsichtlich aktueller Rechtschreibnormen angepasst.)

3.2. Pilotierungen

Nach Abschluss der Entwicklungsphase wurden die Schreibaufgaben in mehreren Wellen pilotiert. Insgesamt wurden von 2007 bis 2010 vier Studien durchgeführt, in deren Rahmen (auch) Schreibaufgaben erprobt wurden.

---

[10] In der in Kapitel 6 präsentierten Teilstudie wird ausführlich auf die theoretischen und empirischen Aspekte eingegangen, welche für oder gegen ein textmusterspezifisches (respektive gegen oder für ein textmusterübergreifendes) Konstrukt von *Schreibkompetenz* sprechen.

[11] Versuchsweise wurden auch solche Aufgabentypen, beispielsweise zur Korrektur, zur Überarbeitung, zur Textgliederung oder zum Einsatz von Kohäsionsmitteln, in geringer Anzahl entwickelt und erprobt. Dabei offenbarten sich jedoch zahlreiche methodische Probleme. So zeigte sich zum Teil eine zu geringe Varianz in den Schülerleistungen oder die Antworten wiesen keine hinreichende Konstruktpassung (Trennschärfe, Modell-Fit) auf. Darüber hinaus erwiesen sich die gewonnenen Informationen als speziell und auf einen einzelnen Aspekt oder wenige Einzelaspekte beschränkt, als dass diese einen essentiellen Beitrag zur Ermittlung der Schreibkompetenz als Ganzes liefern konnten. Gerade unter dem benötigten erheblichen zeitlichen Aufwand, welchen auch die Bearbeitung dieser Aufgabentypen erfordert, musste der Einsatz solcher Aufgaben verworfen werden.

Zwei dieser Studien wurden in allen deutschen Bundesländern durchgeführt, eine in 8 der 16 Länder, eine ausschließlich in Berlin und Brandenburg. Dabei waren alle Schulformen des allgemeinbildenden Schulsystems der jeweils teilnehmenden Länder beteiligt. Jede Aufgabe wurde an 300 bis 600 Schülerinnen und Schülern erprobt. Zwei der vier Studien wurden in den Jahrgangsstufen 8 bis 10 durchgeführt, eine ausschließlich in Jahrgangsstufe 8, eine in den Jahrgangsstufen 9 und 10.

3.3. Auswertungsschemata

Die Auswertungsschemata wurden in einem wechselseitigen Prozess der Erprobung der Aufgabe und Anpassung der Schemata anhand der Pilotierungsresultate entwickelt. Dabei wurden zwei verschiedene Auswertungssysteme ausgearbeitet, ein holistisches und ein analytisches (vgl. Kapitel 2.3.2).

Im Rahmen des analytischen Systems wird eine Beurteilung der Schülertexte mittels mehrerer dichotomer Kriterien (17–18 pro Aufgabe) vorgenommen. Bei diesen Kriterien handelt es sich um Erfüllungsmerkmale, hierunter beispielsweise die Kriterien *Orthografie* (Merkmal: mehr/weniger als 4 % der Wörter weisen orthografische Fehler auf), *Grammatik* (Merkmal: mehr/weniger als 2 % der Wörter weisen grammatikalische Fehler auf), *Textsorte* (Merkmal: zutreffende Textsorte – ja/nein) oder *Perspektive* (einheitliche und angemessene Perspektive – ja/nein). Tabelle 3.3.1 gibt einen aufgabenübergreifenden Überblick über die Kriterien.

Tabelle 3.3.1: ***Analytische Kriterien zur Beurteilung der Schreibaufgaben.***

Formale Kriterien	**Sprachliche Kriterien**	**Strukturelle Kriterien**	**Inhaltliche Kriterien**
Schriftbild	Orthografie	Textsorte	*aufgabenspezifisches Inhaltskriterium 1*
Auswertbarkeit	Grammatik	ggf. Textelemente	*aufgabenspezifisches Inhaltskriterium 2*
Textumfang	Zeichensetzung	Perspektive	*aufgabenspezifisches Inhaltskriterium 3*
	Wortschatz (Korrektheit)	Arrangement der Inhalte	*aufgabenspezifisches Inhaltskriterium 4*
	Sprachlicher Stil	Formale Strukturierung	*(ggf. aufgabenspez. Inhaltskriterium 5)°*

° Für einige Aufgaben wurden vier, für einige fünf inhaltliche Anforderungskriterien definiert.

Das holistische Auswertungssystem dient einer ganzheitlichen Beurteilung von Schülertexten (vgl. Kapitel 2.3.2.). In der konkreten Realisierung für die Normierungsstudie besteht das System aus vier verschiedenen Skalen: einer Globalskala zur Erfassung der Qualität des Textes insgesamt sowie drei Subskalen zu den Dimensionen *Inhalt*, *Stil* und *sprachliche Richtigkeit*. Auf diesen Subskalen wird der jeweilige Teilaspekt ganzheitlich, aber von den anderen Teilaspekten getrennt, beurteilt. Bei den vorliegenden Skalen handelt es sich um vier- bis fünfstufige Ordinalskalen, welche eine graduelle Abstufung in der Ausprägung der Merkmale zulassen.

Bei der Entwicklung des holistischen Systems wurde sich stark an den von NAEP (*National Assessment of Educational Progress)* entwickelten Skalen zur ganzheitlichen Textbeurteilung orientiert. Die eingesetzten Globalskalen sind angepasste, übersetzte Versionen der entsprechenden NAEP-Skalen (NAEP, 2001, 2011b). Das System (mit anderen, d. h. altersentsprechenden Stufenbeschreibungen und Skalenabschnitten) wurde bereits erfolgreich in den Schreibstudien des IQB im Bereich *Deutsch* der Primarstufe eingesetzt (Böhme et al., 2009).[12]

[12] Im Rahmen einer durch das IQB durchgeführten Vorstudie zur empirischen Erprobung der Skalen erwies sich für die Sekundarstufe I eine fünfstufige Globalskala als geeigneter als eine sechsstufige, wie sie in NAEP sowie in der Primarstufe im Rahmen der IQB-Studien, zum Einsatz kamen.

Die Globalskalen sind hierbei textmusterspezifisch, innerhalb der Textmuster jedoch aufgabenübergreifend identisch. Bei den inhaltlichen Subskalen handelt es sich um aufgabenspezifische Textbeurteilungsinstrumente. Die Stilskalen bestehen aus einem textmusterspezifischen Gerüst und enthalten aufgabenspezifische Füllstellen (Slots); so ist beispielsweise der textsortenspezifische Tempusgebrauch ein allgemeines Stilmerkmal für alle informierenden Texte, welches Tempus jedoch zu wählen ist, wird aufgabenspezifisch festgelegt, so etwa das Präteritum für einen Zeitungsbericht, das Präsens für eine Bauanleitung. Die Skala zur sprachlichen Richtigkeit ist aufgaben- und textmusterunabhängig und somit für alle Aufgaben in einer einheitlichen Version gültig. Im Anhang finden sich unter A.3.3.1. bis A.3.3.7 die drei Globalskalen, die drei Gerüste der Stilskalen sowie die Sprachskala. Ergänzend finden sich unter A.3.3.8 bis A.3.3.11 für die beiden Beispielaufgaben die spezifischen Inhalts- und Stilskalen.

Alle holistischen Skalen wurden dabei um mehrere prototypische Beispieltexte, sogenannte *Benchmark*-Texte, ergänzt.

3.4. Normierungsstudie: Datenerhebung

Die Datenerhebung zur Normierung der Aufgaben fand im April und Mai 2011 statt. Insgesamt nahmen 2996 Schülerinnen und Schüler der neunten und zehnten Jahrgangstufe aus insgesamt 278 Schulen aus allen deutschen Bundesländern teil. Das Durchschnittsalter der Teilnehmenden lag bei 15 Jahren und 11 Monaten (*SD* = 0.80 Jahre); 50.9 % der Teilnehmenden waren Mädchen; 87.5 % gaben als Herkunftssprache Deutsch an. Die Stichprobenziehung, welche durch das Data Processing and Research Center (DPC) in Hamburg vorgenommen wurde, beruhte auf einem vom DPC entwickelten Algorithmus, welcher bei der Auswahl der Schulen und Klassen Schulformen und Bundesländer berücksichtigte und eine diesbezüglich repräsentative Teilnahme anstrebte.

Im Rahmen der Studie wurden zwölf[13] freie Schreibaufgaben eingesetzt, darunter vier argumentierende, vier informierende, davon je zwei beschreibende sowie zwei berichtende, und vier narrative Aufgaben. Jede Schülerin und jeder Schüler bearbeitete zwei der zwölf

[13] Von den insgesamt 36 entwickelten Aufgaben wurden 12 Aufgaben gewählt, die in den Pilotierungen eine hinreichende Qualität und Funktionalität für Large-Scale-Erfassungen zeigten und sich in der dafür vorgesehenen Bearbeitungszeit von 20 Minuten pro Aufgabe seitens der Schülerinnen und Schüler als in hinreichendem Umfang bearbeitbar erwiesen.

Aufgaben. Jede Aufgabe wurde von circa 500 Schülerinnen und Schülern bearbeitet. Die Aufgaben waren gemäß dem zugrunde liegenden Testdesign im Rahmen einer Spiral-Verknüpfung über insgesamt 24 Testhefte verteilt, so dass jede Aufgabe mit jeder Aufgabe – zumindest mittelbar – verbunden war. Tabelle 3.4.1. illustriert das Prinzip der Spiral-Verknüpfung exemplarisch.

Tabelle 3.4.1: ***Spiral-Verknüpfung von vier Aufgaben mit je zwei Aufgaben pro Testheft.***

Testheft	**Aufgabenzuordnung**	
Testheft 01	Aufgabe 1	Aufgabe 2
Testheft 02	Aufgabe 2	Aufgabe 3
Testheft 03	Aufgabe 3	Aufgabe 4
Testheft 04	Aufgabe 4	Aufgabe 1

3.5. Kodierung

Die Kodierung der freien Schreibaufgaben erfolgte im Zeitraum von August 2011 bis Mai 2012 in drei Wellen, je eine Welle für jedes Textmuster. Für jede Welle und jedes Kodiersystem (analytisch vs. holistisch) wurden jeweils 8 bis 9 Kodiererinnen und Kodierer akquiriert. Bei den Kodierenden handelte es sich vorwiegend um Studierende des Lehramts *Deutsch* oder einem verwandten Fach (bspw. Linguistik, Germanistik).

Die jeweiligen Kodierenden wurden mehrfach im Beurteilen der Schülertexte anhand der Auswertungsschemata geschult. Das Prozedere war hierbei wie folgt:

Zunächst wurden den Kodierenden die Aufgaben und die Auswertungsschemata vorgestellt und erläutert. Anschließend erhielten die Kodierenden ein Paket von 20–40 Schülertexten pro Aufgabe, welche sie beurteilen sollten. Alle Kodierenden erhielten dieselben Texte. Im Rahmen der folgenden Schulung wurden diejenigen Texte, welche die Kodierenden uneinheitlich beurteilt hatten, gemeinsam besprochen. Pro Welle und Zweig fand dieser Zyklus an Textbeurteilung und gemeinsamer Besprechung zwei- bis dreimal statt.

Nach der zweiten bzw. dritten Schulung erwiesen sich die Kodierergebnisse im Rahmen des Testsamples als hinreichend reliabel.[14] Anschließend wurde die Gesamtmenge der Schülertexte gleichmäßig auf die Kodierenden aufgeteilt mit einem gewissen Prozentsatz an Texten (ca. ein Drittel), der zur Reliabilitätsbestimmung von jeweils zwei Kodierenden beurteilt wurde. Dabei wurden für alle Texte Globalurteile erfasst, für die holistischen Subskalen sowie die analytischen Kriterien wurde eine Zufallsauswahl von zwei Drittel bis drei Viertel der aufgabenspezifischen Textmenge herangezogen.

Die Interraterreliabilitäten für die finalen Kodierungen der Gesamtmenge der Schülertexte erwiesen sich als weitgehend zufriedenstellend. In den Tabellen 3.5.1 bis 3.5.3 sind die Reliabilitäten sowie die prozentualen Übereinstimmungen abgebildet. Während Reliabilitäten lediglich Auskunft über die Beurteilungskonsistenz geben, indem die Rangreihen miteinander vergleichen werden, gibt das Interrater-Agreement, dessen einfachste und geläufigste Form die Form der prozentuale Übereinstimmung ist, den Anteil der exakten Matches an (LeBreton & Senter, 2008; Tinsley & Weiss, 2000; Wirtz & Caspar, 2002). Reliabilitätsmaße hingegen tragen schiefen Verteilungen Rechnung: Wird beispielweise das Kriterium *zutreffende Perspektive* von 90 bis 95 % der Schülerinnen und Schüler erfüllt, schlagen die Uneinheitlichkeiten im kritischen 5–10%-Bereich stärker zu Buche als im Rahmen der Bestimmung der prozentualen Übereinstimmung. Böhme et al. (2009) weisen darauf hin, dass bei Beurteilungen auf mehrstufigen Skalen wie den hier eingesetzten Skalen zur holistischen Beurteilung von Schreibaufgaben sehr selten hohe exakte Übereinstimmungen vorliegen.

> „Die Forderung nach einer exakten Übereinstimmung ist in einem solchen Fall aber weder notwendig noch sinnvoll. Erforderlich ist vielmehr, dass alle Rater durch ihre Urteile dieselbe Anordnung der Aufsätze in einer Rangreihe erzeugen und sich hinsichtlich der relativen Qualität der Schülerantworten einig sind. Mitunter wird gefordert, dass sich die Rater bei der Bewertung der Fähigkeitsausprägung um nicht mehr als eine Kategorie der Skala unterscheiden sollte, sodass die Kompetenzen zwar nicht identisch, aber sehr ähnlich eingeschätzt werden.“ (S. 297)

Die Einbeziehung der Nachbarstufe als tolerierbare Abweichung wird als *nähere* oder *relative Übereinstimmung* bezeichnet und wird häufig als das relevante Maß der Urteilerübereinstimmung für Ratings anhand mehrstufiger Ratingskalen angegeben (Lehmann, 1987; A. Neumann, 2007). Die Beachtung der näheren Übereinstimmung erweist sich gerade im Rahmen von Aufsatzbeurteilungen, bei welchen mehrere verschiedene Kriterien beachtet werden müssen, als sinnvoll. Die Mehrheit der Schülertexte entsprechen nicht einer prototypischen Realisierung einer bestimmten Stufe, oftmals unterscheiden sich die

[14] Die Maßstäbe zur Interpretation der Güte der Urteilerübereinstimmung werden im Fortfolgenden erläutert.

prototypischen Stufenzuordnungen über die einzelnen Merkmale eines Textes; so kann in einem Schülertext beispielweise ein sehr elaborierter Wortschatz (prototypisch für die stilistische Stufe 4) mit einer uneinheitlichen Perspektive (prototypisch für die stilistische Stufe 2 oder geringer) kombiniert sein. Ein Großteil der Schülertexte ist somit aufgrund der Vielfalt und Nichtprototypizität nicht zweifelsfrei auf einer Stufe, sondern im Bereich zwischen zwei Stufen zu verorten.

Für die holistische Kodierung der Schülertexte im Rahmen der Normierung liegen die exakten prozentualen Übereinstimmung bei 50 bis 60 % für die Globalskalen sowie die Subskalen *Inhalt*, *Stil* und *sprachliche Richtigkeit*, die näheren oder relativen Übereinstimmungen bei über 96 % (vgl. Tabelle 3.5.1). Es zeigen sich keine nennenswerten Abweichungen nach Textmuster oder Aufgabe (siehe Angaben in Klammer in Tabelle 3.5.1.).

Tabelle 3.5.1: ***Mittleres Interrater-Agreement für die holistischen Skalen.***

Skala	**exakte prozentuale Übereinstimmung**	**nähere prozentuale Übereinstimmung**
Global	52 % (48–58 %)	96 % (94–98 %)
Inhalt	60% (51–66 %)	98% (94–100 %)
Stil	57% (47–64 %)	98% (96–99 %)
Sprachliche Richtigkeit	59% (48–65 %)	98% (96–100 %)

In Klammern: Spektrum der aufgabenspezifischen Übereinstimmungen; exakte prozentuale Übereinstimmung: Anteil der Urteile, bei welchen beide Pseudokodierer dieselbe Stufe vergaben; nähere prozentuale Übereinstimmung: Anteil der Urteile, bei welchen beide Pseudokodierer Stufen mit einer maximalen numerischen Abweichung von 1 vergaben.

Zur Bestimmung der Interrater-Reliabilität wurden Intraklassenkorrelationen (kurz: ICC) berechnet. In der Literatur finden sich verschiedene Richtwerte zur Interpretation der Werte bzw. Wertebereiche. Für die holistische Textbeurteilung wurde sich an der Interpretation nach Cicchetti und Prusoff (1983) bzw. Cicchetti und Sparrow (1981) orientiert, nach welcher Werte <.40 als gering/schlecht, Werte zwischen .40 und .59 als mäßig/ausreichend, Werte zwischen .60 und .74 als gut und Werte ≥ .75 als ausgezeichnet/exzellent zu bewerten sind

(Stieglitz 2008). In bisherigen Schreibassessment-Studien (wie u. a. DESI) lagen für die Beurteilung freier Schreibaufgaben weitestgehend Werte im moderaten bis guten Bereich vor (Böhme et al., 2009; A. Neumann, 2007). Tabelle 3.5.2 gibt die mittleren ICCs für die vier holistischen Skalen an.

Tabelle 3.5.2: ***Mittlere Interraterreliabilität für die holistischen Skalen.***

Skala	ICC
Global	.64
Inhalt	.66
Stil	.55
Sprachliche Richtigkeit	.67

Alle Skalen erweisen sich als zufriedenstellend reliabel. Die nur als moderat zu interpretierende Übereinstimmung der stilistischen Urteile lässt sich darauf zurückführen, dass besonders bei der Beurteilung dieser Dimension die Raterinnen und Rater nichtprototypische Schülertexte nicht gemäß einer interindividuell einheitlichen Tendenz, kodiert hatten, vielmehr wurde gehäuft gleichsam ‚über Kreuz' kodiert; dies bedeutet, dass etwa für zwei vorliegende Texte im Grenzbereich zwischen Stufe 2 und 3 der eine Kodierer den ersten Text der Stufe 2, den zweiten der Stufe 3 zuordnete, der zweite Kodierer den ersten Text der Stufe 3, den zweiten der Stufe 2 zuordnete; hinsichtlich exakter und vor allem näherer prozentualen Übereinstimmung weist die Stilskala keine schlechteren Ergebnisse als die anderen Skalen auf.

Für die dichotomen analytischen Variablen wurde als Reliabilitätsmaß Cohen's Kappa bestimmt (Fleiss & Cohen, 1973). Auch hier finden sich in der Literatur Schwankungen, wie die einzelnen Wertebereiche zu interpretieren sind; in Anlehnung an Landis und Koch (1977) und Shrout (1998) wurden für die Reliabilitätstestung der Beurteilung im Rahmen des analytischen Auswertungsschemas Werte < .40 als zu gering/schwach, Werte zwischen .40 und .50 als mäßig, Werte zwischen .50 und .60 als zufriedenstellend und Werte >.60 als gut beurteilt. Darüber hinaus wurden wiederum prozentuale Übereinstimmungen erfasst. Tabelle 3.5.3 zeigt die entsprechenden Ergebnisse.

Mit Ausnahme der Variable *Wortschatz (Korrektheit)*, erweisen sich alle Variablen als mindestens in akzeptablem Maße reliabel. Darüber hinaus zeigen sich die Variablen

sprachlicher Stil und *Arrangement der Inhalte* als problematisch, hier liegen bei nur mäßigen Reliabilitäten sehr niedrige prozentuale Übereinstimmungen vor. *Textsorte* weist ebenfalls eine nur mäßige Reliabilität auf; hier liegt jedoch eine hinreichende prozentuale Übereinstimmung vor; die Abweichung des Reliabilitätsmaßes rührt hier von der Verteilungsschiefe des Kriteriums (die überwiegende Mehrheit der Texte erfüllt das Kriterium *Textsorte*).

Tabelle 3.5.3: ***Mittlere Interraterreliabilität und mittleres Interrater-Agreement für die analytischen Kriterien.***

Kriterium	Kappa	prozentuale Übereinstimmung
Orthografie	.67	86
Grammatik	.50	80
Zeichensetzung	.63	82
Wortschatz (Korrektheit)	.33	70
Sprachlicher Stil	.43	73
Textsorte	.44	85
Textelemente	.84	93
Perspektive	.59	86
Arrangement der Inhalte	.43	72
Formale Strukturierung	.73	88
*Inhaltsvariablen**	*.64*	*84*

* Inhaltsvariablen sind aufgabenspezifisch und somit nicht über Aufgaben hinweg identisch, hier handelt es sich um eine nicht auf kriteriale Identität beruhende Zusammenfassung aller Inhaltsvariablen aller Aufgaben.

3.6. Skalierung

Zur Modellierung von *Schreibkompetenz* wurden in einem ersten Schritt zur Ermittlung der Aufgabenschwierigkeiten und Stufengrenzen[15] die Globalurteile der einzelnen Texte herangezogen und in einem mehr-, d. h. dreidimensionalen ordinalen Raschmodell mit der Software *ConQuest* (Version 2.0) (Wu, Adams & Wilson, 1998) skaliert. Als Globalwerte

[15] Stufengrenzen sind diejenigen Skalenwerte, bei welchen es gleich wahrscheinlich ist, einen Bewertung X (oder niedriger zu erhalten) zu erhalten wie eine Bewertung X+1 (oder höher).

wurden alle Skalenstufen von 1 bis 5 einbezogen. Mit Stufe 0 (= zu kurz für eine sinnvolle Beurteilung) bewertete Texte sowie Texte, die mit *missing by intention* kodiert wurden, wurden nicht in diese Analyse einbezogen. Vorabanalysen zeigten, dass die Stufe 0 sich nicht im Sinne einer weiteren Stufe unterhalb der Stufe 1 einordnen lässt. Bei entsprechenden IRT-Modellierungen mit separater Stufe 0 verortete diese sich breiter als andere Stufen mit Peak zwischen den Stufen 1 und 2.

Die Modellierung erfolgte hierbei textmusterspezifisch, d. h. die Aufgaben eines Textmusters wurden jeweils einer Dimension zugeordnet. Gründe für die mehrdimensionale Modellierung waren sowohl theoretische als auch empirische. Eine psychologisch und fachdidaktisch valide Beschreibung ließe sich textmusterübergreifend nur sehr abstrakt und vage realisieren, um sinnvoll interpretier- und anwendbar zu sein, oder konstituierte sich aus einer Liste an bedingten Disjunktionen. Aus empirischer Sicht erwies sich das dreidimensionale Modell hinsichtlich der Modellpassung leicht besser als das eindimensionale. In Kapitel 6 werden sowohl dieser Punkt als auch andere empirische Aspekte für oder gegen eine textmuster-spezifische Modellierung näher erläutert, untersucht und diskutiert.

Zur Ermittlung der Personenfähigkeiten wurden in einem zweiten Schritt die Schüler-antworten, die mit Stufe 0 bewertet wurden oder als *missing by intention* kategorisiert wurden, in die Analyse miteinbezogen. Dazu erfolgte eine Umkodierung dieser Werte auf Stufe 1. Dieses Vorgehen entspricht dem Prozedere in anderen Kompetenzbereichen, nicht interpretierbare, unvollständige und zu kurze Antworten als niedrigsten Ausprägungswert zu re-klassifizieren. Dieser so gewonnene Datensatz wurde erneut wie oben beschrieben skaliert, jedoch wurden die Aufgabenschwierigkeiten und Schwellen (Stufengrenzen) auf die Werte aus der ersten Modellierung fixiert. Zur Schätzung der Personenfähigkeiten wurde die Jahrgangstufe (9. vs. 10.) durch die Aufnahme der entsprechenden Variablen in das Hintergrundmodell berücksichtigt.

Abbildung 3.6.1 illustriert die Ergebnisse der Skalierung hinsichtlich der Lage der Stufen-grenzen sowie die textmusterspezifischen Aufgabenverteilungen. Es zeigten sich für alle Aufgaben sinnvoll geordnete Stufengrenzen sowie annähernde Normalverteilungen für die latenten Werte.

Für alle Aufgaben zeigten sich zufriedenstellende Trennschärfen (0.78–0.88) und Fitwerte (Infit/weighted MNSQ[16]: 0.98–1.15).

[16] *MNSQ*: Meansquare

Abbildung 3.6.1: ***Wright Map: Latente Verteilungen und Thresholds für die Schreibaufgaben – Textmuster als verschiedene Dimensionen.***

	Dimension					
	Argumentation		Information		Narration	
Logits	Verteilung	Stufengrenzen (Thresholds)	Verteilung	Stufengrenzen (Thresholds)	Verteilung	Stufengrenzen (Thresholds)
5						
4						
	X			$I3_{4/5}$		
3	X	$A2_{4/5}$		$I1_{4/5}$ $I2_{4/5}$		
	X		X			
	X	$A1_{4/5}$ $A3_{4/5}$	X			
	XX		X	$I4_{4/5}$		$N3_{4/5}$ $N4_{4/5}$
	XX		XX		X	
2	XX	$A4_{4/5}$	XX		X	$N1_{4/5}$
	XXX		XXX		XX	$N2_{4/5}$
	XXXX		XXXX		XX	
	XXXX		XXX		XXX	
	XXXXX		XXXXX		XXXX	
1	XXXXXX		XXXXXX		XXXXXX	
	XXXXXXX		XXXXXXX		XXXXXXX	
	XXXXXX		XXXXXXXX	$I2_{3/4}$	XXXXXXXX	
	XXXXXXXX		XXXXXXXX	$I1_{3/4}$	XXXXXXXXX	$N3_{3/4}$
	XXXXXXX		XXXXXXXXX	$I3_{3/4}$ $I4_{3/4}$	XXXXXXXXX	$N4_{3/4}$
	XXXXXXXX	$A2_{3/4}$	XXXXXXXX		XXXXXXXXX	$N2_{3/4}$
0	XXXXXXX	$A1_{3/4}$ $A3_{3/4}$	XXXXXXXXX		XXXXXXXXX	$N1_{3/4}$
	XXXXXXXX	$A4_{3/4}$	XXXXXXXX		XXXXXXXXX	
	XXXXXXX		XXXXXXXXX		XXXXXXXXX	
	XXXXXX		XXXXXXXX		XXXXXXXXX	
	XXXXXX		XXXXXXXX		XXXXXXXXX	
-1	XXXXXXX		XXXXXX		XXXXXXXX	$N3_{2/3}$
	XXXXXXX		XXXXXX		XXXXXXX	
	XXXXX	$A2_{2/3}$	XXXXX	$I1_{2/3}$ $I2_{2/3}$ $I3_{2/3}$	XXXXXX	$N1_{2/3}$ $N2_{2/3}$
	XXXXX		XXXX	$I4_{2/3}$	XXXXX	$N4_{2/3}$
	XXXX	$A1_{2/3}$	XXXX		XXXX	
	XXXX		XXXX		XXXX	
-2	XXX		XX		XX	
	XXX	$A3_{2/3}$	XX		X	
	X	$A4_{2/3}$	X		X	$N1_{1/2}$ $N2_{1/2}$ $N3_{1/2}$
	X	$A2_{1/2}$	XX		X	
	X		X		X	
-3	X		X	$I1_{1/2}$ $I2_{1/2}$ $I4_{1/2}$		$N4_{1/2}$
	X			$I3_{1/2}$		
	X		X		X	
	X					
		$A3_{1/2}$	X			
-4		$A1_{1/2}$				
		$A4_{1/2}$				
-5						

A1-A4: argumentierende Aufgaben; I1-I4: informierende Aufgaben; N1-N4: narrative Aufgaben; Indizierung: Stufengrenze zwischen den jeweils angegebenen Stufen (z. B. $_{1/2}$: Stufengrenze zwischen Stufe 1 und Stufe 2). X repräsentiert 25.2 Personen.

Für weitere Analysen wurden die WLEs (*Weighted Likelihood Estimates*) (Warm, 1989) sowie fünf PVs (*Plausible Values*) (Mislevy, Beaton, Kaplan & Sheehan, 1992) als Schätzungen der Personenfähigkeiten bestimmt.

Die ermittelten Leistungs- und Schwierigkeitswerte wurden nun auf einen Mittelwert von 500 und eine Standardabweichung von 100 normiert. Die 500-100-Skala hat sich im Rahmen bildungsbezogener Kompetenzmessung etabliert (Klieme et al., 2006; Köller, Knigge & Tesch, 2010; A. Neumann, 2007; Prenzel, 2007; Stanat, Pant, Böhme & Richter, 2012). Referenzgruppe für die Standardisierung der Skala bildeten alle Schülerinnen und Schüler der Jahrgangsstufe 9 allgemeinbildender Schulen der Bundesrepublik Deutschland. Um von der Stichprobe auf die Gesamtheit aller Schülerinnen und Schüler zu schließen, wurden den Schülerinnen und Schülern Populationsgewichte zugeordnet, welche die Faktoren *Schulform* und *Bundesland* berücksichtigen. Die Ermittlung der Gewichte erfolgte durch das DPC in Hamburg.

3.7. Standard-Setting

Aufgrund der Skalierung konnten sowohl die Schwierigkeiten der Schreibaufgaben bzw. deren einzelne Stufen als auch die Personenfähigkeiten auf einer einheitlichen kontinuierlichen Skala (pro Textmuster) verortet werden, die aufgrund der entsprechenden Modellierung als Skala der textmusterspezifischen Schreibkompetenz interpretiert werden kann. Undefiniert blieb bisher jedoch, wie bestimmte Ausprägungen auf dieser Skala inhaltlich-qualitativ und auch normativ, etwa bezüglich der Erreichung von Regelstandards gemessen an den Bildungsstandards, interpretiert werden sollten.

Zu diesem Zwecke wurde ein Standard-Setting-Verfahren durchgeführt. Unter *Standard-Setting* versteht man eine Methode zur Definition von Stufen und/oder Stufengrenzen (Cizek & Bunch, 2007; Pant, Tiffin-Richards & Köller, 2010).

> „Idealtypisch wird in einem Standard-Setting Verfahren ein Panel aus Expert/innen konstituiert, das in einem iterativen Verfahren aus Einzelurteilen und Gruppendiskussionen zur Festlegung von Cut-Scores kommt.“ (Pant et al., 2010, S. 176)

Als Panel wurden insgesamt 18 Expertinnen und Experten aus den Bereichen der Fachdidaktik, der empirischen Bildungsforschung und Bildungsadministration sowie der Schulpraxis (Lehrkräfte) aus insgesamt elf deutschen Bundesländern sowie (je eine Person) aus Österreich und aus der Schweiz akquiriert.

Insgesamt wurden drei Standard-Setting-Verfahren – je eines pro Textmuster – durchgeführt. Für jedes Standard-Setting-Verfahren bestand das Panel aus 12 bis 14 Personen.

Für das Standard-Setting im Bereich *Schreiben* wurde sich an dem entsprechenden Verfahren bei NAEP (2011a) orientiert. Dem Verfahren lag eine sogenannte *Body-of-Work*-Methode zugrunde (Kingston, Kahl, Sweeney & Bay, 2001); dies bedeutet, dass alle am Standard-Setting Beteiligten für mehrere Personen (hier Schülerinnen/Schüler) je einen Satz an Aufgabenbearbeitungen der jeweiligen Person erhalten und diesen Aufgabensatz als Ganzes einer von mehreren qualitativen Kategorien zuordnen (Hambleton & Pitoniak, 2006; Details im Fortfolgenden).

In Vorbereitung auf das Standard-Setting wurden durch die fachdidaktische Projektbetreuung unter der Leitung von Prof. Dr. Albert Bremerich-Vos und unter psychometrischer Beratung des IQB vorläufige Stufenbeschreibungen der Kompetenzstufenmodelle entwickelt.[17] Diese Beschreibungen dienten als qualitative Einstufungskriterien für die Beurteiler.

Ein Standard-Setting-Verfahren erstreckte sich über drei Experten-Beurteilungsrunden. Tabelle 3.7.1. bietet einen Überblick über die einzelnen Schritte für das Standard-Setting *Schreiben.* Die einzelnen Schritte werden im Fortfolgenden detailliert erläutert.

[17] Zur Genese dieser Beschreibungen: vgl. das Folgekapitel (3.8.).

Tabelle 3.7.1: ***Überblick: Standard-Setting im Kompetenzbereich Schreiben.***

	Gruppe 1	**Gruppe 2**	**Gruppe 3**
Runde 1	individuelle Bewertung von $25_{\text{Gruppe 1}}$ geordneten Schülertextpaaren	individuelle Bewertung von $25_{\text{Gruppe 2}}$ geordneten Schülertextpaaren	individuelle Bewertung von $25_{\text{Gruppe 3}}$ geordneten Schülertextpaaren
Feedback zu Runde 1	Übereinstimmungs-tabelle$_{\text{Gruppe 1}}$	Übereinstimmungs-tabelle$_{\text{Gruppe 2}}$	Übereinstimmungs-tabelle$_{\text{Gruppe 3}}$
	Impact Daten (Schülerverteilungen) gemäß idealen Cutpoints$_{\text{alle}}$		
Runde 2	Diskussion der Schüler-beurteilungen$_{\text{Gruppe 1}}$	Diskussion der Schüler-beurteilungen$_{\text{Gruppe 2}}$	Diskussion der Schüler-beurteilungen$_{\text{Gruppe 3}}$
	erneute individuelle Bewertung der $25_{\text{Gruppe 1}}$ Schülertextpaare	erneute individuelle Bewertung der $25_{\text{Gruppe 2}}$ Schülertextpaare	erneute individuelle Bewertung der $25_{\text{Gruppe 3}}$ Schülertextpaare
	Ermittlung aller individuellen idealen Cutpoints.		
	Für jede Stufengrenze Textauswahl in und um das Spektrum der individuellen Cutpoints.		
Runde 3 (Feinjustierung)	individuelle Bewertung von 8–10 ungeordneten Schülertextpaaren$_{\text{alle}}$ für die Stufengrenze 1-2		
	individuelle Bewertung von 8–10 ungeordneten Schülertextpaaren$_{\text{alle}}$ für die Stufengrenze 2-3		
	individuelle Bewertung von 8–10 ungeordneten Schülertextpaaren$_{\text{alle}}$ für die Stufengrenze 3-4		
	Individuelle Bewertung von 8–10 ungeordneten Schülertextpaaren$_{\text{alle}}$ für die Stufengrenze 4-5		
	Ermittlung der finalen Stufengrenzen		
	Prüfung auf und ggf. Durchführung von Glättungen, Anpassungen		

Zu Runde 1 wurden die Panel-Mitglieder in drei verschiedene Gruppen à vier bis fünf Personen aufgeteilt. Jede Gruppe bestand aus einer Fachdidaktikerin oder einem Fachdidaktiker, ein bis zwei Lehrkräften sowie ein bis zwei Vertreterinnen bzw. Vertreter der Bildungsforschung und/oder Bildungsadministration.

Jedes Mitglied einer Gruppe erhielt nun ein geordnetes Booklet an Schülertextpaaren, d. h. die jeweils beiden Aufgabenbearbeitungen einer Schülerin oder eines Schülers. Das Booklet bestand aus insgesamt 25 Schülertextpaaren. Diese waren gemäß den aus der Skalierung gewonnenen WLE-Werten der Schülerinnen und Schüler aufsteigend (von sehr niedrigem

WLE bis sehr hohem WLE) geordnet. Zu Runde 1 erhielten alle Gruppenmitglieder dasselbe Booklet; verschiedene Gruppen erhielten verschiedene Booklets. Die Panelmitglieder sollten nun die Schülertextpaare bzw. die entsprechende Schülerin / den entsprechenden Schüler anhand der vorläufigen Kompetenzstufenbeschreibungen einer solchen Kompetenzstufe zuordnen. Diese Zuordnung erfolgte individuell und ohne Absprache und Kommunikation mit anderen Gruppen- oder Panelmitgliedern. Die Panelmitglieder wurden darüber informiert, dass das Booklet gemäß (bisher ermittelter) Schülerfähigkeit geordnet ist; ihnen wurde jedoch mitgeteilt, dass sie dennoch frei beurteilen können und nicht strikt aufsteigend bewerten müssen, die Ordnung diene nur zur Orientierung am Leistungsspektrum und der Information, in welchem Bereich dieses Spektrums man sich gerade bewegt. Auch wurde den Mitgliedern freigestellt, in welcher Reihenfolge sie die Texte lesen und beurteilen, ein Vor- und Zurückspringen war uneingeschränkt möglich. Für die Zuordnung der Schülertextpaare zu den Stufenbeschreibungen hatten die Panelmitglieder ca. fünf bis sechs Stunden Zeit.

Im Anschluss an Runde 1 wurden die Ergebnisse durch Mitarbeiter des IQB zusammengetragen, die Beurteilungen, i. e. die Stufenzuordnungen der Schülertextpaare, aller Mitglieder einer Gruppe wurden in einer Tabelle spaltenweise gegenübergestellt. Diese Gegenüberstellung wurde den Gruppenmitgliedern am Folgetag zur Eröffnung von Runde 2 ausgehändigt. Zusätzlich wurden anhand der Ergebnisse aus Runde 1 mittels logistischen Regressionen die nach den derzeitigen Beurteilungen idealen Stufengrenzen berechnet und die daraus resultierenden Schülerverteilungen bestimmt. Diese Impact-Daten wurden den Panelmitgliedern ebenfalls zu Eröffnung der zweiten Runde präsentiert.

Mit der Übersicht, bei welchen Schülertextpaaren gruppeninterne Uneinigkeit bestand, und mit dem Wissen über die Verteilungskonsequenzen (Impact-Daten), sollten die einzelnen Gruppen jeweils gruppenintern über die Schülertextpaare und ihre Einstufung diskutieren. Explizites Ziel dieser Diskussion war es nicht, einen Konsens zu erreichen, vielmehr ging es um einen Austausch der Einschätzungen der Schülerkompetenzen und der Beachtung, Gewichtung und Bewertung einzelner Aspekte. Es sollte somit die Möglichkeit geboten werden, auf bestimmte individuelle Über- oder Unterbewertungen aufmerksam zu werden und diese ggf. zu korrigieren sowie eventuelle Interpretations- und somit auch Bewertungsambiguitäten aufzudecken. Aus diesem Grund sollten die Gruppenmitglieder idealtypischerweise nicht nur über die uneinheitlich bewerteten, sondern über alle Schülertextpaare diskutieren und sich austauschen. Für die Diskussionsrunde standen rund 3 Stunden Zeit zur Verfügung.

Im Anschluss an die Diskussionsrunde sollten die Panelmitglieder dieselben 25 Schülertextpaare (wie in Runde 1) erneut individuell einstufen, diesmal ggf. unter Modifikation ihres Bewertungsverhaltens auf Basis der Erkenntnisse der Diskussionsrunde. Zur Beurteilung standen ca. 3 Stunden Zeit zur Verfügung.

Im Anschluss an Runde 2 wurden auf Basis der Expertenbeurteilungen die rechnerisch idealen Stufengrenzen (*Cutpoints*) ermittelt. Diese wurden erneut mittels Berechnung logistischer Regressionen bestimmt. Hierzu wurden zunächst alle Bewertungen hinsichtlich aller vier Stufengrenzen dichotomisiert (1 vs. 2–5; 1–2 vs. 3–5; 1–3 vs. 4–5; 1–4 vs. 5). Die so gewonnenen dichotomen Urteile für jede Stufengrenze wurden als abhängige Variable für die Berechnung des jeweils idealen Cutpoints herangezogen. Als unabhängige Variable diente der dem Schülertextpaar jeweils zugeordnete WLE des Schülers bzw. der Schülerin. Auf diesem Wege wurde für jedes Panelmitglied der ideale Cutpoint für alle vier Stufengrenzen berechnet.

In Vorbereitung auf Runde 3 wurden diese individuellen idealen Cutpoints für jede Stufengrenze in eine aufsteigende Reihenfolge gebracht. Nun wurden für jeden Cutpoint 8 bis 10 Schülertextpaare nach folgenden Kriterien ausgewählt: Es wurde mindestens ein Schülertextpaar herangezogen, welches gemäß der Skalierung einem WLE minimal unterhalb des niedrigsten individuellen idealen Cutpoint zugeordnet war, sowie mindestens ein Schülertextpaar, welches einem WLE minimal oberhalb des höchsten individuellen idealen Cutpoint zugeordnet war; die anderen Textpaare wurden aus dem Intervall zwischen dem niedrigsten und dem höchsten individuellen Cutpoint gezogen. Tabelle 3.7.2 verdeutlicht die Textauswahl an einem Beispiel.

Tabelle 3.7.2: ***Individuelle ideale Cutpoints aller Panelmitglieder für eine Stufengrenze und darauf basierende Textauswahl für Feinjustierungsrunde (hier Textmuster: argumentieren).***

Geordnete individuelle Cutpoints für Stufengrenze 1-2 gemäß der Einstufungen in Runde 2	Auswahl an Schülertextpaaren (WLEs - geordnet)
	275
	275
277	
277	
277	
277	
	279
	279
280	
280	
280	
280	
309	
309	
338	338
338	338
338	
	348
	358

Geordnete individuelle Cutpoint: rechnerisch ermittelte Stufengrenzen pro Standard-Setting-Teilnehmer. Auswahl an Schülertextpaaren: WLE-Zuordnungen (gemäß Skalierung) real vorliegender Schülertextpaare; Spektrum der ausgewählten Schülertextpaare soll das Spektrum der individuellen Cutpoints abdecken sowie einen überschaubaren Skalenbereich darunter und darüber.

Die Panelmitglieder erhielten nun für jeden Cutpoint ein Booklet von in beschriebener Weise erwählten 8 bis 10 Schülertextpaaren, diesmal in unsortierter, randomisierter Abfolge. Die Mitglieder wurden über die Nichtsortierung informiert und instruiert, die Schülertextpaare eines Booklets jeweils einer der beiden Kompetenzstufen zuzuordnen, die unmittelbar unterhalb oder oberhalb des zu ermittelnden Cutpoints liegen.

Diese Methode diente der Feinjustierung des zu ermittelnden Cutpoints. Die Auswahl mindestens eines Schülertextpaares unterhalb des Spektrums der individuellen Stufengrenzen

und mindestens eines Schülertextpaares oberhalb dieses Spektrums, sollte sicherstellen, dass jedes Panelmitglied auch bei zu den Vorrunden konsistentem Beurteilungsverhalten beide relevanten Kompetenzstufen vergeben konnte. Zusätzlich wurden die Beurteilenden darüber informiert, dass sie nicht ausgewogen bewerten müssen und es durchaus der Fall sein kann, dass sie beispielsweise siebenmal den einen Wert und nur einmal den anderen Wert vergeben. So sollte vermieden werden, dass die Beurteilenden aufgrund des neuen ggf. asymmetrischen Samples eine implizite Verschiebung in ihrem Beurteilungsverhalten vornehmen.

Die Feinjustierung (Runde 3) wurde als individuelle ‚Hausaufgabe' seitens der Panelmitglieder geleistet. Ihnen stand hierfür eine Woche Zeit zur Verfügung.

Im Anschluss an Runde 3 wurden erneut mittels logistischer Regressionen die idealen individuellen Cutpoints bestimmt. Tabelle 3.7.3 verdeutlicht die Entwicklungen der Verteilungen der individuellen Cutpoints über die drei Bewertungsrunden eines Standard-Settings.

Tabelle 3.7.3: ***Individuelle ideale Cutpoints aller Panelmitglieder für eine Stufengrenze für alle drei Runden eines Standard-Settings (hier Textmuster: narrativ).***

Geordnete individuelle Cutpoints für Stufengrenze 3-4 nach Runde 1	Geordnete individuelle Cutpoints für Stufengrenze 3-4 nach Runde 2	Geordnete individuelle Cutpoints für Stufengrenze 3-4 nach Runde 3
494	546	552
541	546	555
543	546	555
545	546	562
546	546	570
548	548	571
560	548	571
560	584	576
579	599	585
594	599	585
600	599	585
651	600	585
663	600	586

Der Mittelwert der individuellen Cutpoints aus Runde 3 diente als vorläufiger Cutpoint für die jeweilige Stufe.

In einem letzten Arbeitsschritt zur Ermittlung der Stufengrenzen wurde nun geprüft, in welchem Spektrum um den ermittelten vorläufigen Cutpoint keine realen Schülerdaten vorlagen. Dieses Spektrum ist der Bereich, in welchem der ermittelte Cutpoint auf Basis der dem Standard-Setting zugrunde liegenden Daten nicht differenziert, d. h. eine Verschiebung des Cutpoints in diesem Bereich führt zu keiner Veränderung der Schülerverteilungen. Anhand dieser Cutpointbereiche wurde nun geprüft, ob es rechnerisch möglich ist, die Cutpoints innerhalb der Bereiche so zu setzen, dass pro Modell identische Stufenbreiten für die Nichtrandstufen 2 bis 4 vorliegen. Diese Möglichkeit war für alle Stufengrenzen und für alle drei Textmuster gegeben, die Setzung der idealisierten Cutpoints wurde dementsprechend vorgenommen. In Tabelle 3.7.4 sind die finalen Stufengrenzen und -breiten dargestellt.

Tabelle 3.7.4: ***Finale Stufengrenzen und -breiten der Kompetenzstufen für die textmusterspezifischen Kompetenzstufenmodelle im Bereich Schreiben.***

Textmuster	Stufe 1	Stufe 2	Stufe 3	Stufe 4	Stufe 5
argumentieren	bis 369	370 bis 464	465 bis 559	560 bis 654	ab 655
informieren	bis 382	383 bis 473	474 bis 564	565 bis 655	ab 656
narrativ	bis 411	412 bis 492	493 bis 573	574 bis 654	ab 655

3.8. Kompetenzstufenbeschreibungen

Die Kompetenzstufenbeschreibungen wurden in zwei Schritten generiert. Der erste Schritt war hierbei dem Standard-Setting vorgelagert und diente der Entwicklung von Stufenbeschreibungen, welche für das Standard-Setting-Verfahren einsetzbar sind.

Für die Genese dieser Beschreibungen wurde wie folgt vorgegangen: Aufgrund einer hinreichend ähnlichen Aufgabenschwierigkeit aller Aufgaben eines Textmusters (vgl. Abbildung 3.6.1) konnten die Globalstufen (auf Textqualitätsebene) als Indikatoren für die Kompetenzstufen (auf Schülerfähigkeitsebene) herangezogen werden. Für jede Globalstufe aller Aufgaben wurden nun die zugehörigen mittleren Werte für die Subskalen *Inhalt*, *Stil* und *sprachliche Richtigkeit* sowie für die analytischen Kriterien bestimmt. Anhand dieser Zuordnung konnte bestimmt werden, in welchem Maße die entsprechenden Kriterien auf der jeweiligen Stufe erfüllt sind. Tabelle 3.8.1 stellt diese Verhältniszuordnung am Beispiel einer

Aufgabe dar. Diese Verhältnisse wurden im Anschluss unter Beteiligung fachdidaktischer und psychometrischer Expertinnen und Experten verbalisiert.[18]

Tabelle 3.8.1: ***Mittelwerte für die analytischen Kriterien und die holistischen Subskalen nach Globalstufen anhand einer Beispielaufgabe.***

Skala/Kriterium	Stufe 1	Stufe 2	Stufe 3	Stufe 4	Stufe 5
Holistisch - Inhalt	1.10	1.95	2.70	3.33	3.95
Holistisch - Stil	1.30	2.10	2.66	3.06	3.35
Holistisch - Sprache	2.05	2.27	2.52	2.90	3.15
Orthografie	0.33	0.49	0.57	0.86	0.95
Grammatik	0.39	0.59	0.76	0.82	0.90
Zeichensetzung	0.22	0.39	0.51	0.67	0.90
Sprachlicher Stil	0.06	0.27	0.50	0.77	0.80
Textsorte	0.39	0.61	0.95	0.97	1.00
Perspektive	0.22	0.59	0.91	0.96	1.00
Arrangement der Inhalte	0.17	0.32	0.59	0.62	0.90
Formale Strukturierung	0.28	0.54	0.65	0.78	0.90
Inhalt1	0.11	0.27	0.58	0.73	0.90
Inhalt2	0.11	0.41	0.71	0.79	0.80
Inhalt3	0.50	0.66	0.92	0.96	1.00
Inhalt4	0.06	0.46	0.76	0.83	0.85

Die auf diesem Wege entwickelten vorläufigen Kompetenzstufenbeschreibungen unterlagen dem Anspruch, inhaltlich keine Veränderungen zu erfahren, da die Zuordnungen im Standard-Setting auf diesen Beschreibungen fußten.

[18] Die Hauptaufgabe bestand hierbei darin, den prozentualen Erfülltheitsquoten passende sprachliche Modifikatoren und Quantifikatoren zuzuordnen und diese einheitlich zu verwenden. So wurde etwa die Quote 0.97 für das Kriterium *Textsorte* wie folgt verbalisiert: *Der Textsorte wird nahezu immer entsprochen*.

Im Standard-Setting-Verfahren hatten die Expertinnen und Experten jedoch die Möglichkeit, Rückmeldung zu den Stufenbeschreibungen zu geben, ob bestimmte Passagen ambig, missverständlich oder zu vage sind. Im Standard-Setting-Verfahren selbst bestand die Möglichkeit durch Rückfragen diese Aspekte zu vereindeutigen bzw. zu konkretisieren, sodass das Standard-Setting-Verfahren nicht durch eventuelle Miss- und Fehlinterpretationen beeinflusst wurde.

Im Anschluss an das Standard-Setting wurden im Rahmen eines Workshops mit teilnehmenden Expertinnen und Experten aus Fachdidaktik und Psychometrie die Stufenbeschreibungen anhand der mündlichen und schriftlichen Rückmeldungen der Standard-Setting-Teilnehmenden sprachlich überarbeitet und optimiert. Die so gewonnenen Verbalisierungen konstituieren die finalen Kompetenzstufenbeschreibungen in den Kompetenzstufenmodellen für den Bereich *Schreiben*. Die Stufenbeschreibungen der drei Modelle finden sich im Anhang unter A.3.8.1 bis A.3.8.3.

Im Rahmen der Kompetenzstufenmodelle werden diese so beschriebenen Kompetenzstufen schließlich normativ als Aussagen über die graduelle Erreichung der durch die Bildungsstandards definierten erwarteten Schülerkompetenzen interpretiert.

> „Hierfür wurden die folgenden Festlegungen getroffen:
>
> - *Mindeststandards* beziehen sich auf ein definiertes Minimum an Kompetenzen, das alle Schülerinnen und Schüler bis zu einem bestimmten Bildungsabschnitt erreicht haben sollten. Diese unterschreiten die in den Publikationen der KMK festgelegten Kompetenzerwartungen der Regelstandards. Sie beschreiben jedoch ein Kompetenzniveau am Ende der Sekundarstufe I, von dem angenommen werden kann, dass sich Schülerinnen und Schüler, die dieses erreichen, bei entsprechender Unterstützung erfolgreich in die berufliche Erstausbildung integrieren werden.
> - *Regelstandards* beziehen sich auf Kompetenzen, die im Durchschnitt von den Schülerinnen und Schülern bis zu einem bestimmten Bildungsabschnitt erreicht werden sollen und den von der KMK definierten Kompetenzzielen entsprechen.
> - Als *Regelstandard plus* wird ein Leistungsbereich definiert, der über den Regelstandards liegt und als Zielperspektive für die Weiterentwicklung von Unterricht angesehen werden kann.
> - *Optimalstandards* beziehen sich auf Leistungserwartungen, die bei sehr guten oder ausgezeichneten individuellen Lernvoraussetzungen und der Bereitstellung besonders günstiger Lerngelegenheiten innerhalb und außerhalb der Schule erreicht werden können und die bei weitem die Erwartungen der KMK-Bildungsstandards übertreffen.“ (IQB, 2014, S. 4).

Die Kompetenzstufenmodelle wurden schließlich im März 2014 von der Kultusministerkonferenz (KMK) beschlossen.

4. Hauptbefunde der Normierungsstudie

In diesem Kapitel werden einige zentrale Ergebnisse der Normierungsstudie im Kompetenzbereich *Schreiben* vorgestellt. Die Darstellung erfolgt hierbei rein deskriptiv und dient als Überblick über die Hauptbefunde der Normierungsstudie. Insofern andere als die in Kapitel 3 erläuterten Analysen durchgeführt wurden, findet sich zu Beginn des jeweiligen Subkapitels eine Darstellung der entsprechenden Analysen. In Kapitel 4.1. werden die Kompetenzstufenverteilungen für die drei textmusterspezifischen Kompetenzstufenmodelle dargelegt. Kapitel 4.2. befasst sich mit gruppenspezifischen Unterschieden von Schreibkompetenzen. In Kapitel 4.3. werden Ergebnisse zu inhaltlichen, stilistischen und orthografisch-grammatischen Schreibfähigkeiten präsentiert. In Kapitel 4.4. erfolgt schließlich eine Zusammenfassung und Einordnung der Ergebnisse.

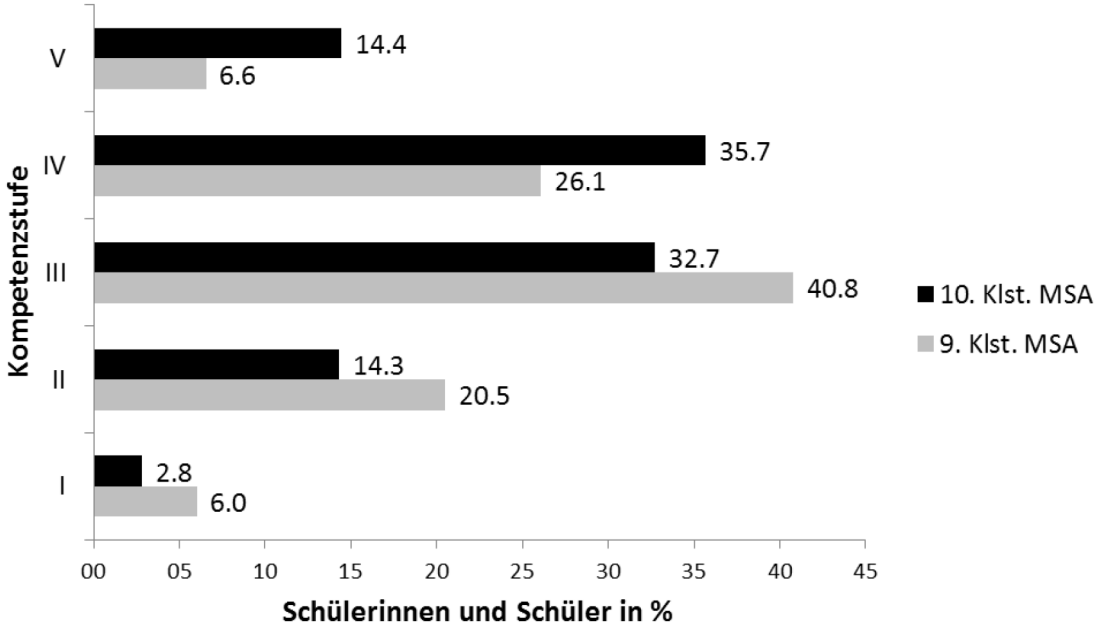
4.1. Stufenverteilungen für die Kompetenzstufenmodelle im Bereich *Schreiben*

Für den Bereich Schreiben wurden, wie in Kapitel 3 erläutert, drei Kompetenzstufenmodelle entwickelt, je eines für argumentierende, informierende und narrative Texte. Die Kompetenzstufenmodelle beziehen sich auf den Mittleren Schulabschluss, weshalb in nachfolgenden Kompetenzstufenverteilungen auch nur diejenigen Schülerinnen und Schüler betrachtet wurden, welche den MSA anstreben.

Die Abbildungen 4.1.1 bis 4.1.3 zeigen die Kompetenzstufenverteilungen für alle Schülerinnen und Schüler, die den MSA anstreben,[19] für die 9. und 10. Jahrgangsstufe.

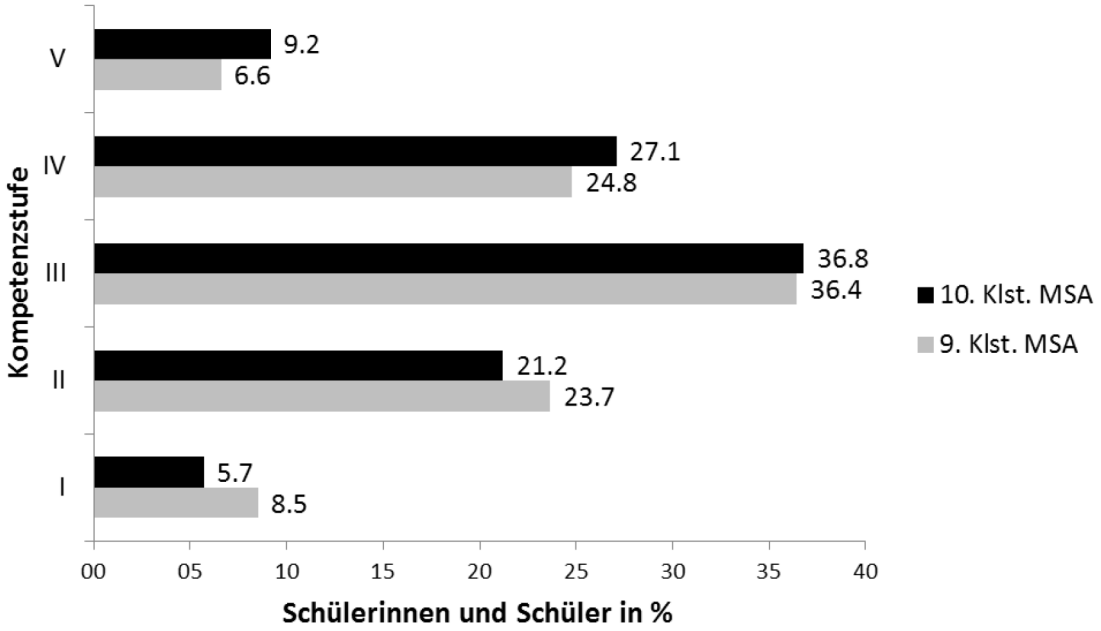
[19] inkl. Schülerinnen und Schülern, die nach dem MSA einen weiteren Abschluss im allgemeinbildenden Schulsystem (Abitur) anstreben

Abbildung 4.1.1: ***Verteilung der Schülerinnen und Schüler der 9. und 10. Klassenstufe, die den MSA anstreben, auf die Kompetenzstufen des Kompetenzstufenmodells Schreiben für argumentierende Texte.***



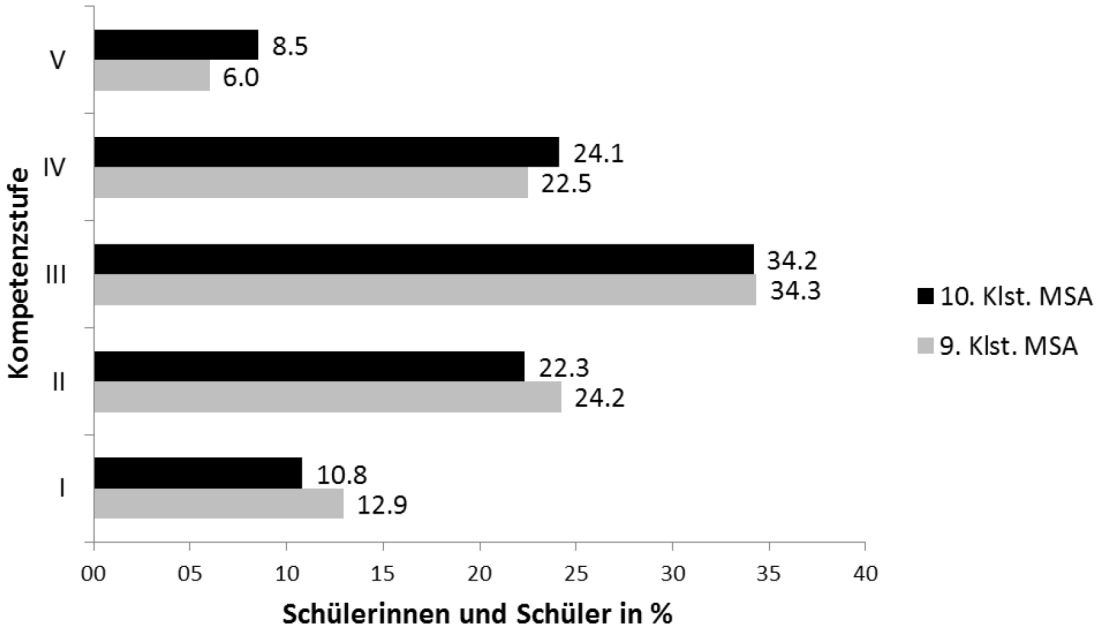
Für das Schreiben argumentierender Texte erreichen in der 9. Jahrgangsstufe 6 Prozent der Schülerschaft noch nicht dem MSA-relevanten Mindeststandard, in der 10. Jahrgangsstufe verfehlen diesen nur noch knapp 3 Prozent. Circa ein Fünftel der Schülerschaft erreicht in Klassenstufe 9 noch nicht den durch die Bildungsstandards vorgegebenen Regelstandard, in Klassenstufe 10 verringert sich dieser Anteil auf knapp 15 Prozent. Über vier Fünftel der Schülerinnen und Schüler erreichen somit in der relevanten Jahrgangsstufe die standardisierten Vorgaben.

Abbildung 4.1.2: ***Verteilung der Schülerinnen und Schüler der 9. und 10. Klassenstufe, die den MSA anstreben, auf die Kompetenzstufen des Kompetenzstufenmodells Schreiben für informierende Texte.***



Für das Schreiben informierenderTexte erreichen in der 9. Jahrgangsstufe rund 9 Prozent der Schülerschaft noch nicht den MSA-relevanten Mindeststandard, in der 10. Jahrgangsstufe verfehlen diesen nur noch knapp 6 Prozent. Annähernd ein Viertel der Schülerschaft erreicht in Klassenstufe 9 noch nicht den durch die Bildungsstandards vorgegebenen Regelstandard, in Klassenstufe 10 verringert sich dieser Anteil auf rund 21 Prozent. Knapp drei Viertel der Schülerinnen und Schüler erreichen somit in der relevanten Jahrgangsstufe die standardisierten Vorgaben.

Abbildung 4.1.3: ***Verteilung der Schülerinnen und Schüler der 9. und 10. Klassenstufe, die den MSA anstreben, auf die Kompetenzstufen des Kompetenzstufenmodells Schreiben für narrative Texte.***



Für das Schreiben narrativer Texte erreicht in der 9. Jahrgangsstufe ein Achtel der Schülerschaft noch nicht den MSA-relevanten Mindeststandard, in der 10. Jahrgangsstufe reduziert sich dieser Anteil auf circa ein Zehntel. Annähernd ein Viertel der Schülerschaft erreicht in Klassenstufe 9 noch nicht den durch die Bildungsstandards vorgegebenen Regelstandard, in Klassenstufe 10 verringert sich dieser Anteil auf rund 22 Prozent. Gut zwei Drittel der Schülerinnen und Schüler erreichen somit in der relevanten Jahrgangsstufe die standardisierten Vorgaben.

Bei der Betrachtung der Schülerverteilungen zeigen sich zunächst zwei Auffälligkeiten. Zum einen scheinen die Schülerinnen und Schüler im argumentierenden Schreiben besser abzuschneiden als im informierenden Schreiben als im narrativen Schreiben. Zum anderen scheint der Lernzuwachs[20] von Klassenstufe 9 zu Klassenstufe 10 für argumentierendes Schreiben höher zu sein als für informierendes und narratives Schreiben.

Dabei erscheint ein Vergleich absoluter Ausprägungen verschiedener Kompetenzen jedoch als inhaltlich wenig sinnvoll, da ein gemeinsamer intrinsischer Bezugsrahmen (bspw. eine

[20] Es sei darauf hingewiesen, dass die Unterschiede zwischen den beiden Klassenstufen nicht direkt als „Lernzuwachs“ interpretierbar sind. Die Studie erfolgte quer-, nicht längsschnittlich. Die Interpretation der Jahrgangsunterschiede als Lernzuwächse ist demnach nur unter weitgehendem Ausschluss von Kohorteneffekten gestützt. Kohortenspezifische Einflüsse bei einem Unterschied von lediglich einer Jahrgangsstufe sind jedoch als eher als gering einzuschätzen.

gemeinsame Skala) fehlt. Setzt man jedoch die Interpretation der Kompetenzstufen als Mindeststandards, Regelstandards etc. als einheitlichen Bezugsrahmen voraus, lassen sich die Verteilungen durchaus sinnvoll miteinander vergleichen. Eine Aussage über diesen normativen Interpretationsrahmen hinaus ist jedoch nicht möglich.

Mittels Mann-Whitney-U-Tests ermittelte Vergleiche der Verteilungen bestätigen einen Verteilungsunterschied zwischen den drei Textmustern mit bedeutsam positiverer Verteilung für argumentierendes Schreiben als für informierendes Schreiben als für narratives Schreiben für beide Jahrgangsstufen (alle $p < .05$).

Hinsichtlich der vermutenden Zuwächse von Jahrgangsstufe 9 zu 10 sei auf das Folgekapitel verwiesen, in welchem diese Unterschiede anhand der metrischen Leistungsdaten untersucht werden und somit der Varianzen innerhalb der Stufen Rechnung getragen wird.

4.2. Gruppenspezifische Unterschiede von Schreibkompetenzen

Im Folgenden werden Unterschiede von Schülergruppen dargestellt in Abhängigkeit von den Faktoren *Klassenstufe*, *Geschlecht*, *Sprachhintergrund* und *Schulform*. Folgende Ausprägungen lagen für die einzelnen Variablen vor:

Klassenstufe:

- 9. Klassenstufe
- 10. Klassenstufe

Geschlecht:

- männlich
- weiblich

Sprachhintergrund:

- Deutsch als Herkunftssprache
- Deutsch nicht als Herkunftssprache / andere Herkunftssprache als Deutsch

Schulform:

- Hauptschule
- Schulformen mit mehreren Bildungsgängen (Regelschule, Mittelschule, Sekundarschule)

- Integrierte Gesamtschule
- Realschule
- Gymnasien

Klassenstufe und *Schulform* sind Variablen, deren Ausprägungen extern festgelegt sind und zur Testung bekannt waren. *Geschlecht* und *Sprachhintergrund* wurde von den Schülerinnen und Schülern erfragt; *Sprachhintergrund* wurde hierbei durch folgende Fragestellung erfasst: *Welche Sprache hast du in deiner Familie zuerst gelernt (Erstsprache/Muttersprache)?*[21]

Eine Betrachtung der Ergebnisse nach Sprachhintergrund ist vor allem aus dem Grunde interessant, dass es sich bei *Schreiben Deutsch* um eine Kompetenz der sprachlichen Ausdrucksfähigkeit im Rahmen des konzeptionellen Muttersprachunterrichts handelt, und im Rahmen der Betrachtungen somit Gruppen, die dieser Konzeption gerecht werden, mit solchen, die dieser Konzeption nicht entsprechen, verglichen werden.

Die schulformbezogenen Analysen betrachten vor allem den institutionellen Rahmen, in welchem Schreiben curricular vermittelt wird. Damit verbunden sind auch schulform-spezifisch Vorbereitungen auf unterschiedliche Abschlüsse, für deren Erreichen auch andere Bildungsstandards definiert sind (vgl. Kapitel 2.1.4.).

Die genannten Variablen sind teilweise nicht unabhängig voneinander. Alle 6 Kreuztabellen inklusive Chi-Quadrat-Statistik finden sich im Anhang unter A.4.2.1 bis A.4.2.6.

Während *Klassenstufe*, *Geschlecht* und *Sprachhintergrund* jeweils unabhängig voneinander sind (vgl. Tabellen A.4.2.2, A.4.2.3 & A.4.2.5), zeigen sich Abhängigkeiten aller Variablen zu *Schulform*.

Klassenstufe weist insofern einen Zusammenhang mit *Schulform* auf, als dass in der 10. Klassenstufe kaum noch Schülerinnen und Schüler vertreten sind, die eine Hauptschule besuchen (vgl. Tabelle A.4.2.1).

Der Zusammenhang zwischen *Geschlecht* und *Schulform* liegt darin begründet, dass der Mädchenanteil auf Realschulen und Gymnasien höher ist als in Hauptschulen, Schulen mit mehreren Bildungsgängen und Gesamtschulen (vgl. Tabelle A.4.2.4).

[21] Ergänzend wurde der Hinweise gegeben: *Bitte nur ein Kästchen ankreuzen.* Dabei wurde eine Liste von Einzelsprachen im Ankreuzformat vorgegeben sowie die Kategorie *eine andere Sprache, und zwar* gefolgt von einem Freitextfeld.

Ein Zusammenhang zwischen *Sprachhintergrund* und *Schulform* besteht insofern, dass der Anteil der Schülerinnen und Schüler nichtdeutscher Herkunftssprachen in Hauptschulen mit einem knappen Viertel deutlich höher liegt als in Realschulen (ca. 15 %) als in Gymnasien (knapp 7 %) (vgl. Tabelle A.4.2.6).

Diese Zusammenhänge in der Stichprobe entsprechen weitgehend den Zusammenhängen in der Gesamtheit aller Schülerinnen und Schüler in der Bundesrepublik (Statistisches Bundesamt, 2011).

Für die im Folgenden vergleichenden Analysen wurde folgendes Vorgehen gewählt: Zunächst wurden die aufgabenspezifischen Leistungswerte der Schülerinnen und Schüler skaliert. Die drei Textmuster wurden als unabhängige Dimensionen modelliert. Im Unterschied zu dem in Kapitel 3.6. beschriebenen Vorgehen wurden zur Schätzung der Schülerleistungswerte neben der Variable *Klassenstufe* auch die Variablen *Geschlecht*, *Sprachhintergrund* und *Schulform* mit in das Hintergrundmodell aufgenommen. Die geschätzten Leistungswerte[22] wurden anschließend auf eine Metrik mit einem Mittelwert von 500 und einer Standardabweichung von 100 standardisiert; Referenzgruppe für die Standardisierung waren alle Neuntklässler des allgemeinbildenden Schulsystems der Bundesrepublik Deutschland.[23]

Die Analysen wurden wie folgt durchgeführt: Alle Effekte wurden mittels einfaktoriellen Varianzanalysen berechnet.[24] Zusätzlich wurden alle Effekte auch isoliert betrachtet (einfaktorielle Varianzanalysen, t-Tests). Für die Variable *Schulform*, die mehr als zwei Ausprägungen aufweist, wurden zusätzlich Tests zur Gruppenhomogenität berechnet.

Effektstärken für Mittelwertsunterschiede wurden als Cohen's *d* bestimmt. Effektstärkenvergleiche wurden mittels Berechnung von *Z*-Werten, d. h. standardisierte Differenzen zwischen den entsprechenden Fisher-*z*-Werten der Effektstärken, durchgeführt (J. Cohen, 1988; Howell, 2002).

---

[22] Als Schülerleistungswerte dienten fünf *Plausible Values* (vgl. Kapitel 3.6.).
[23] vgl. Kapitel 3.6.
[24] Im Rahmen von Vorabanalysen wurde geprüft, ob Interaktionen vorliegen, welche eine sinnvolle Interpretation der Haupteffekte einschränken könnten und ob sich die Haupteffekte möglicherweise überlagern. Hierfür wurden zunächst mehrfaktorielle Varianzanalysen (gesättigte Modelle, Typ III) berechnet, um mögliche Interaktionen zu detektieren. Lagen keine Interaktionen vor oder lagen für diese über die parallelen Analysen anhand der Plausible Values keine stabilen Ergebnisse vor, wurden die Varianzanalysen nur für Haupteffekte wiederholt. Diese Analysen entsprechen einem varianzanalytischen Modell des Typs II (Regressionsmodell). Die Aussagekraft von Typ-II-Modellen ist bei Nichtvorliegen von Interaktionen stärker (Langsrud, 2003). Da sich in allen präsentierten Analysen keine stabilen statistisch signifikanten Interaktionen zeigten, wurde dieser Analyseschritt in allen Effektberechnungen durchgeführt. Des Weiteren zeigten sich (mit einer Ausnahme) keine bedeutsamen Unterschiede im Befundmuster zwischen den kontrollierten (mehrfaktorielle Varianzanalysen, Typ II) und den im Text berichteten isolierten (einfaktorielle Varianzanalysen) Effekten.

4.2.1. Klassenstufe

Neben der Hauptanalyse, welche alle Schülerinnen und Schüler der 9. Klassenstufe mit allen Schülerinnen und Schülern der 10. Klassenstufe kontrastiert, wurde die Analyse für das Subsample der Schülerinnen und Schüler durchgeführt, welche den Mittleren Schulabschluss oder das Abitur anstreben. Die Wahl dieser Subgruppe liegt darin begründet, dass die in der Normierung eingesetzten Aufgaben die Bildungsstandards für den Mittleren Schulabschluss überprüfen und das entwickelte Kompetenzstufenmodell ein Modell für den MSA ist.

Tabelle 4.2.1.1: ***Unterschiede in den Schreibkompetenzen zwischen Schülerinnen und Schülern der Klassenstufen 9 und 10.***

	Mittelwert argumentieren (SD; SE)	*p*-Wert Cohen's *d*	Mittelwert informieren (SD; SE)	*p*-Wert Cohen's *d*	Mittelwert narrativ (SD; SE)	*p*-Wert Cohen's *d*
9. Klassenstufe (alle) *n* = 1842	508 (80; 1.9)	< .001 0.46	505 (77; 1.8)	.001 0.19	508 (79; 1.8)	< .001 0.34
10. Klassenstufe (alle) *n* = 1154	544 (73; 2.2)		520 (72; 2.1)		534 (70; 2.1)	
9. Klassenstufe (MSA/Abitur) *n* = 1498	528 (69; 1.8)	< .001 0.31	519 (70; 1.8)	.209 0.05	528 (67; 1.7)	< .001 0.15
10. Klassenstufe (MSA/Abitur) *n* = 1077	550 (70; 2.1)		523 (70; 2.1)		538 (68; 2.1)	

SD: Standardabweichung; SE: Standardfehler des Mittelwerts

Es zeigen sich statistisch bedeutsame Unterschiede in den mittleren Schreibkompetenzausprägungen zwischen den entsprechenden Schülerinnen und Schülern der Klassenstufen 9 und 10.

Betrachtet man nur diejenigen Schülerinnen und Schüler, die den MSA oder das Abitur anstreben, fallen die Effekte kleiner aus, für das informierende Textmuster zeigen sich keine bedeutsamen Unterschiede. Die Unterschiede in dieser Schülergruppe lassen sich als Leistungszuwachs im spezifischen Bereich von Klassenstufe 9 zu Klassenstufe 10 interpretieren. Für *argumentieren* liegt dieser Leistungszuwachs bei etwa 20, für *narrativ* bei rund 10 Punkten. Der Vergleich der Textmuster liefert einen bedeutsamen Unterschied

hinsichtlich der Stärke der Effekte zwischen *argumentieren* einerseits und *informieren* und *narrativ* andererseits ($Z_{argu\text{-}info}$: 4.0; $Z_{argu\text{-}narr}$: 2.5; $Z_{info\text{-}narr}$: 1.5).[25]

4.2.2. Geschlecht

Die Ergebnisse der geschlechtervergleichenden Analyse sind in Tabelle 4.2.2.1 dargestellt. Es zeigt sich ein bedeutsamer Vorteil der Schülerinnen gegenüber den (männlichen) Schülern. Die geschlechtsbezogenen Unterschiede betragen rund 40 Punkte.

Tabelle 4.2.2.1: ***Geschlechtsbezogene Unterschiede in den Schreibkompetenzen zwischen Schülerinnen und Schülern.***

	Mittelwert argumentieren (SD; SE)	*p*-Wert Cohen's *d*	Mittelwert informieren (SD; SE)	*p*-Wert Cohen's *d*	Mittelwert narrativ (SD; SE)	*p*-Wert Cohen's *d*
männlich (alle) *n* = 1471	500 (79; 2.1)	< .001	488 (74; 1.9)	< .001	501 (75; 2.0)	< .001
		0.58		0.61		0.46
weiblich (alle) *n* = 1525	544 (74; 1.9)		533 (71; 1.8)		535 (74; 1.9)	

SD: Standardabweichung; SE: Standardfehler des Mittelwerts

Im Vergleich der Textmuster zeigt sich ein stärkerer Geschlechtereffekt für *argumentieren* und *informieren* als für *narrativ* (beide $Z > 1.96$); *argumentieren* und *informieren* unterscheiden sich nicht bedeutsam voneinander ($Z < 1.96$).

4.2.3. Sprachhintergrund

Für die Analyse nach Sprachhintergrund wurden alle Schülerinnen und Schüler deutscher Herkunftssprache mit Schülerinnen und Schülern anderer Herkunftssprachen kontrastiert.

[25] *Z*-Werte über 1.96 entsprechen einem Signifikanzniveau < .05.

Tabelle 4.2.3.1: ***Unterschiede in den Schreibkompetenzen zwischen Schülerinnen und Schülern deutscher und nichtdeutscher Herkunftssprache.***

	Mittelwert argumentieren (SD; SE)	*p*-Wert Cohen's *d*	Mittelwert informieren (SD; SE)	*p*-Wert Cohen's *d*	Mittelwert narrativ (SD; SE)	*p*-Wert Cohen's *d*
D als HS (alle) *n* = 2621	531 (74; 1.5)	< .001 0.93	517 (74; 1.4)	< .001 0.71	526 (73; 1.4)	< .001 0.90
¬ D als HS (alle) *n* = 374	457 (85; 4.4)		464 (75; 3.9)		459 (74; 3.8)	

SD: Standardabweichung; SE: Standardfehler des Mittelwerts

Wie in Tabelle 4.2.3.1 dargestellt, zeigt sich ein Vorteil der Schülerinnen und Schüler deutscher Herkunftssprache gegenüber Schülerinnen und Schüler nichtdeutscher Herkunftssprache. Der durchschnittliche Unterschied beträgt zwischen rund 50 und 80 Punkten.

Im Vergleich der Textmuster zeigt sich ein stärkerer Effekt für *argumentieren* als für *informieren* ($Z = 2.3$); der Effekt für *narrativ* unterscheidet sich nicht bedeutsam von den anderen (beide $Z < 1.96$).

4.2.4. Schulform

Der Faktor *Schulform* erweist sich in den varianzanalytischen Berechnungen mit $p_{argu} > .001$; $p_{info} = .001$ und $p_{narr} < .001$ als statistisch bedeutsam.

Anschließende paarweise Mittelwertsvergleiche wurden aufgrund der konzeptionellen Abhängigkeit zwischen *Schulform* und *Klassenstufe* nach Klassenstufen getrennt berechnet. Tabelle 4.2.4.1 zeigt die Mittelwerte der Schreibkompetenzwerte für die einzelnen Schulformen getrennt für die Klassenstufen 9 und 10.

Tabelle 4.2.4.1: ***Schulformbezogene Schreibkompetenzen nach Klassenstufe.***

	9. Klassenstufe			10. Klassenstufe		
	argu.	info.	narr.	argu.	info.	narr.
Hauptschulen *n* = 262; 122	426 (69; 4.3)	443 (75; 4.6)	419 (63; 3.9)	478 (60; 5.5)	485 (75; 6.8)	468 (58; 5.2)
MBG *n* = 94; 69	451 (60; 6.2)	492 (68; 7.1)	462 (66; 6.8)	499 (63; 7.6)	501 (78; 9.4)	493 (50; 6.0)
IGS *n* = 203; 134	464 (75; 5.2)	477 (81; 5.7)	468 (69; 4.8)	493 (67; 5.8)	495 (70; 6.1)	493 (61; 5.2)
Realschulen *n* = 528; 345	504 (67; 2.9)	496 (68; 3.0)	508 (59; 2.6)	542 (61; 3.3)	505 (69; 3.7)	525 (59; 3.2)
Gymnasien *n* = 755; 484	559 (56; 2.0)	542 (65; 2.3)	556 (62; 2.2)	584 (61; 2.8)	546 (64; 2.9)	573 (60; 2.7)

Werte: Mittelwerte; Werte in Klammern: Standardabweichung und Standardfehler

In der 9. Klassenstufe zeigt sich für die Textmuster *argumentieren* und *narrativ* folgendes Muster: Schulen mit mehreren Bildungsgängen und Integrierte Gesamtschulen unterscheiden sich nicht bedeutsam. Alle anderen Schulformen weisen signifikante Unterschiede auf (alle $p < .05$). Die Schreibkompetenz der entsprechenden Schülerinnen und Schüler steigt in der Abfolge *Hauptschulen* – {*MBG*; *IGS*} – *Realschulen* – *Gymnasien* an. Der Unterschied zwischen den Extremkategorien, d. h. Hauptschulen und Gymnasien, beträgt rund 140 Punkte ($d = 2.1$). Es zeigen sich keine Unterschiede zwischen den beiden genannten Textmustern (alle $Z < 1.96$).

Für informierendes Schreiben zeigen sich keine Unterschiede zwischen Schulen mit mehreren Bildungsgängen (MBG) und Realschulen. Auch weisen MBG keine bedeutsamen Unterschiede zu den Integrierten Gesamtschulen auf. Alle anderen Unterschiede erweisen sich als statistisch bedeutsam (alle $p < .05$). Die Schreibkompetenz der entsprechenden Schülerinnen und Schüler steigt in der Abfolge *Hauptschulen* – {*MBG*; *IGS*} – {*MBG*; *Realschulen*} – *Gymnasien* an. Der Unterschied zwischen den Extremkategorien, d. h. Hauptschulen und Gymnasien, beträgt rund 100 Punkte ($d = 1.4$) und ist damit weniger stark als für die beiden anderen Textmuster ($Z = 3.6$).

In der 10. Klassenstufe findet sich für *argumentieren* und *narrativ* ein zur Klassenstufe 9 analoges Unterschiedsmuster mit keinen signifikanten Unterschieden zwischen MBG und IGS und statistisch bedeutsamen Unterschieden zwischen allen anderen Kategorien (alle $p < .05$). Die Schreibkompetenz der entsprechenden Schülerinnen und Schüler steigt in der Abfolge

Hauptschulen – {*MBG*; *IGS*} – *Realschulen* – *Gymnasien* an. Der Unterschied zwischen den Extremkategorien, d. h. Hauptschulen und Gymnasien, beträgt rund 110 Punkte (d = 1.8). Es zeigen sich keine Unterschiede zwischen diesen beiden Textmustern (alle $Z < 1.96$).

Für *informieren* finden sich in Klassenstufe 10 keine bedeutsamen Unterschiede zwischen MBG, IGS und Realschulen; diese drei Schulformen bilden eine statistisch homogene Gruppe. Die Schreibkompetenz der entsprechenden Schülerinnen und Schüler steigt in der Abfolge *Hauptschulen* – {*MBG*; *IGS*; *Realschulen*} – *Gymnasien* an. Der Unterschied zwischen den Extremkategorien, d. h. Hauptschulen und Gymnasien, beträgt rund 60 Punkte (d = 0.9) und ist damit weniger stark als in den beiden anderen Textmustern (Z = 5.5).

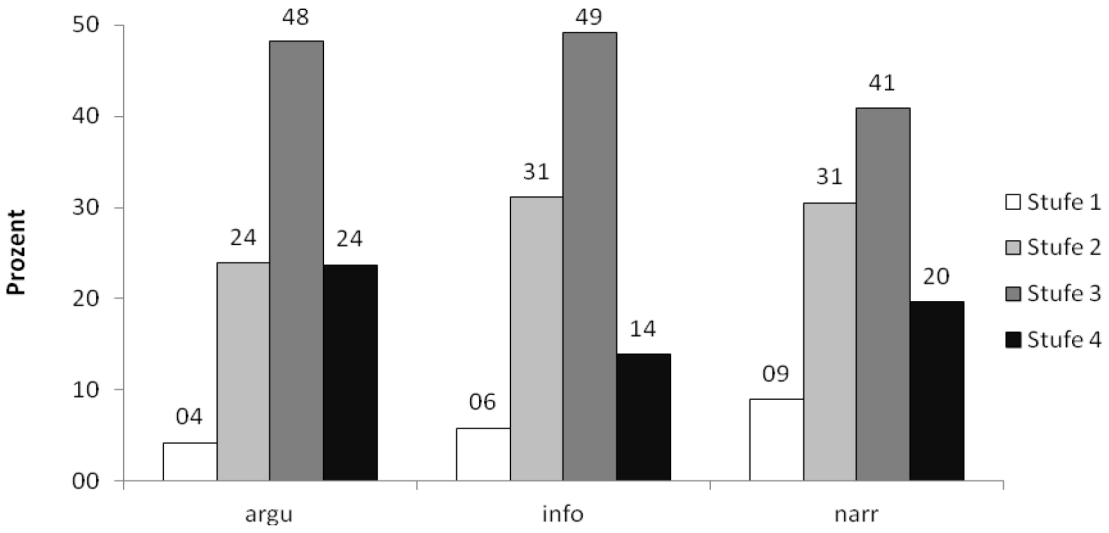
4.3. Ergebnisse bezüglich *Inhalt*, *Stil* und *sprachliche Richtigkeit*

In diesem Kapitel werden die Ergebnisse zu den Schreibkompetenzdimensionen, d. h. der inhaltlichen, stilistischen und orthografisch-grammatischen Schreibkompetenz dargestellt, welche mittels der semiholistischen Subskalen erfasst wurden. In Kapitel 4.3.1. werden die Schülerverteilungen dargestellt, in Kapitel 4.3.2. werden die durchschnittlichen Ausprägungen von Schülergruppen anhand der Merkmale *Klassenstufe*, *Geschlecht*, *Sprachhintergrund* und *Schulform* verglichen.

4.3.1. Schülerverteilungen auf den Subskalen *Inhalt*, *Stil* und *sprachliche Richtigkeit*

Die Abbildungen 4.3.1.1. bis 4.3.1.3. zeigen die Verteilung der Schülerinnen und Schüler gemäß der Bewertung ihrer Texte auf den semiholistischen Subskalen *Inhalt*, *Stil* und *sprachliche Richtigkeit*. Die Skalen für Inhalt und Stil sind so gestaltet, dass sich die jeweilige Stufe 1 als inhaltliche bzw. stilistische Verfehlung der Aufgabenbearbeitung interpretieren lässt (vgl. Anhang A.3.3.4 – A.3.3.6 & A.3.3.8 – A.3.3.11.). Die Skala zur sprachlichen Richtigkeit lässt aufgrund dessen, dass es sich bei orthografischer und grammatischer Richtigkeit um ein streng graduelles Konzept handelt, eine solche Interpretation nicht ohne weitere Zusatzannahmen zu.

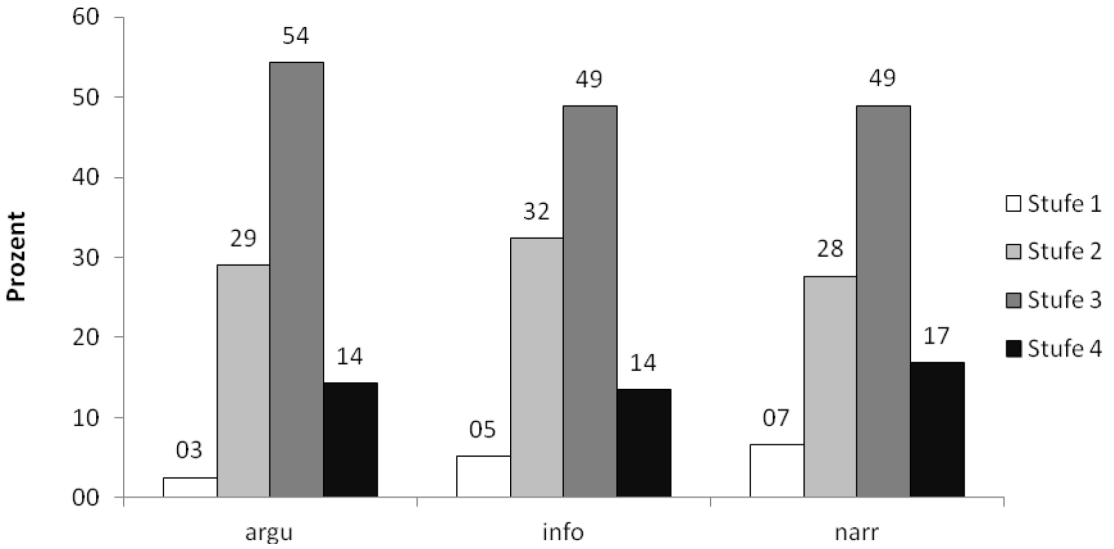
Abbildung 4.3.1.1: ***Schülerverteilung auf der semiholistischen Subskala „Inhalt" nach Textmustern.***



Es zeigt sich, dass je nach Textsorte 4 bis 9 Prozent der Schülerinnen und Schüler im Rahmen ihrer Schreibleistungen das inhaltliche Mindestmaß der Aufgabenbearbeitungen verfehlen. Circa ein Viertel bis ein Drittel erreicht diese Mindestanforderung auf Stufe 2, weitere 40 bis 50 Prozent zeigen Leistungen, die mit der Stufe 3 bewertet wurden, zwischen einem Siebtel und einem Viertel der Schülerschaft erreicht schließlich Stufe 4.

Mittels Mann-Whitney-U-Tests ermittelte Vergleiche der Verteilungen indizieren eine bedeutsam positivere Verteilung für *argumentieren* als für *informieren* als für *narrativ* (alle $p > .05$).

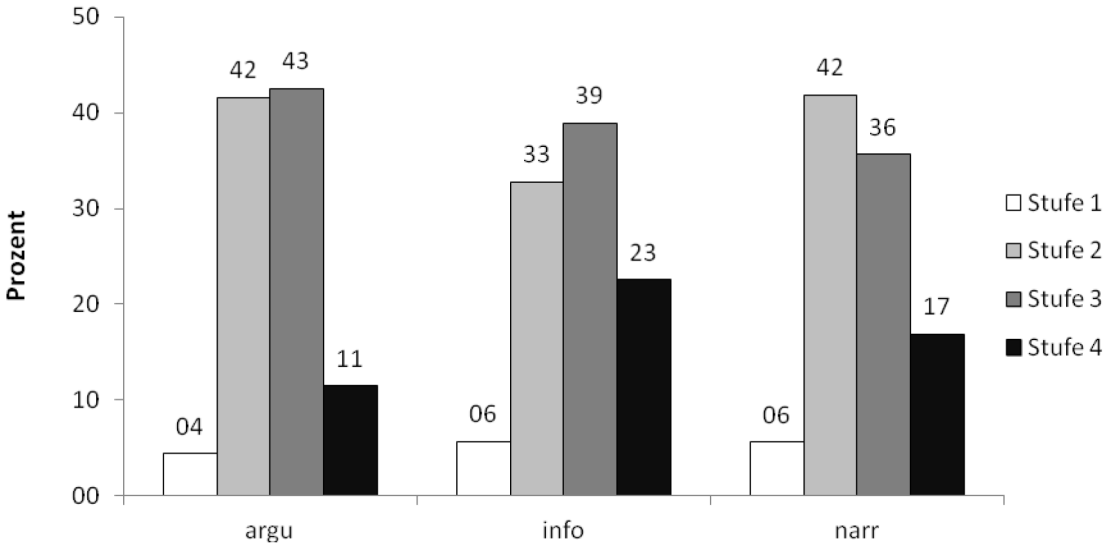
Abbildung 4.3.1.2: ***Schülerverteilung auf der semiholistischen Subskala „Stil“ nach Textmustern.***



Stilistisch verfehlen je nach Textsorte 3 bis 7 Prozent der Schülerinnen und Schüler das erforderliche Mindestmaß; circa 30 Prozent erreichen dieses auf Stufe 2; rund die Hälfte erreicht Stufe 3, etwa 15 Prozent Stufe 4.

Im Vergleich der Verteilungen zeigt sich, dass das argumentierende Textmuster eine positivere Verteilung aufweist als das informierende (p = .007) und das narrative (p = .001), welche sich untereinander nicht unterscheiden (p = .502).

Abbildung 4.3.1.3: ***Schülerverteilung auf der semiholistischen Subskala „sprachliche Richtigkeit“ nach Textmustern.***



Circa 5 Prozent der Schülerinnen und Schüler zeigen Leistungen, die sich der Stufe 1 zuordnen lassen; circa ein Drittel bis zwei Fünftel erreichen jeweils Stufe 2 und 3. Stufe 4 wird schließlich je nach Textmuster von einem Schüleranteil zwischen einem Zehntel und einem Viertel erreicht.

Im Vergleich der Verteilungen zeigt sich, dass das informierende Textmuster eine positivere Verteilung aufweist als das argumentierende und das narrative (beide $p = .001$), welche sich untereinander nicht unterscheiden ($p = .559$).

4.3.2. Gruppenspezifische Unterschiede: *Inhalt*, *Stil* und *sprachliche Richtigkeit*

Um den unterschiedlichen Aufgabenschwierigkeiten Rechnung zu tragen, wurden auch die anhand der Subskalen vergebenen Beurteilungen analog dem in Kapitel 4.2. beschriebenen Vorgehen skaliert. Zur Vergleichbarkeit wurden die Plausible Values auch hier im Anschluss auf einen Mittelwert von 500 und eine Standardabweichung von 100 mit der Referenzgruppe aller Neuntklässler standardisiert. Die so gewonnenen Leistungswerte der Schülerinnen und Schüler lassen sich dadurch als Ausprägungen der Schreibkompetenzdimensionen, d. h. inhaltlicher, stilistischer und orthografisch-grammatischer Schreibkompetenz interpretieren. Das statistische Analysevorgehen war identisch mit dem in Kapitel 4.2. beschriebenen.

4.3.2.1. *Inhalt*

Für die durchschnittliche Ausprägung von inhaltlichen Schreibfähigkeiten zeigen sich Gruppenunterschiede nach *Klassenstufe*, *Geschlecht*, *Sprachhintergrund* und *Schulform*. Die Unterschiede zeigen sich in allen drei Textmustern, teilweise jedoch in unterschiedlicher Stärke.

Der Unterschied hinsichtlich inhaltlicher Schreibkompetenzen zwischen Klassenstufe 9 und 10 beträgt durchschnittlich 20–30 Punkte. Schülerinnen weisen im Schnitt um eine rund 20–30 Punkte erhöhte inhaltliche Schreibkompetenz gegenüber (männlichen) Schülern auf. Schülerinnen und Schüler mit deutscher Herkunftssprache zeigen einen Vorteil von 40–60 Punkten gegenüber Schülerinnen und Schülern mit anderen Herkunftssprachen (vgl. Tabelle 4.3.2.1.1). Es zeigen sich keine bedeutsamen textmusterspezifischen Unterschiede in der Stärke der Effekte (alle $Z < 1.96$).

Tabelle 4.3.2.1.1: ***Durchschnittliche inhaltliche Schreibkompetenzen nach Klassenstufe, Geschlecht und Sprachhintergrund.***

	Mittelwert argumentieren (SD; SE)	*p*-Wert Cohen's *d*	Mittelwert informieren (SD; SE)	*p*-Wert Cohen's *d*	Mittelwert narrativ (SD; SE)	*p*-Wert Cohen's *d*
9. Klassenstufe *n* = 1842	507 (78; 1.8)	< .001	503 (75; 1.7)	< .001	506 (75; 1.7)	< .001
		0.38		0.27		0.36
10. Klassenstufe *n* = 1154	536 (72; 2.1)		523 (73; 2.1)		531 (69; 2.0)	
männlich *n* = 1471	499 (77; 2.0)	< .001	494 (74; 1.9)	< .001	499 (72; 1.9)	< .001
		0.51		0.46		0.48
weiblich *n* = 1525	536 (73; 1.9)		527 (72; 1.8)		534 (71; 1.8)	
D als HS *n* = 2621	526 (73; 1.4)	< .001	518 (73; 1.4)	< .001	521 (73; 1.4)	.002
		0.75		0.58		0.71
¬ D als HS *n* = 374	467 (84; 4.3)		465 (74; 3.8)		479 (73; 3.8)	

SD: Standardabweichung; SE: Standardfehler des Mittelwerts

Ein Vergleich der Schulformen zeigt erneut differente Effekte für die einzelnen Textmuster. Es zeigen sich hierbei keine Unterschiede zwischen den Jahrgangsstufen 9 und 10. Für *argumentieren* und *narrativ* finden sich keine bedeutsamen Unterschiede zwischen *MGB* und *IGS*, alle anderen Schulformen weisen statistisch signifikante Unterschiede auf (alle $p < .05$). Die inhaltliche Schreibkompetenz der entsprechenden Schülerinnen und Schüler steigt in der Abfolge *Hauptschulen – {MBG; IGS} – Realschulen – Gymnasien* an. Für *informieren* weisen die Realschulen keinen bedeutsamen Unterschied zu den MBG und den IGS auf, *MBG* und *IGS* unterscheiden sich jedoch statistisch bedeutsam voneinander ($p = .009$; $d = 0.33$). Alle übrigen paarweisen Vergleiche weisen ebenfalls statistische Signifikanz auf (alle $p < .05$). Die inhaltliche Schreibkompetenz der entsprechenden Schülerinnen und Schüler steigt in der Abfolge *Hauptschulen –{IGS; Realschulen}– {Realschulen, MBG} – Gymnasien* an (vgl. Tabelle 4.3.2.1.2). Die Schulformeffekte unterscheiden sich auch in ihrer Effektstärke nach Textmustern, so weist *argumentieren* einen bedeutsam stärkeren Schulformeffekt auf als *narrativ* als *informieren* (alle $Z > 1.96$).

Tabelle 4.3.2.1.2: ***Schulformbezogene durchschnittliche inhaltliche Schreibkompetenzen nach Klassenstufe.***

	9. Klassenstufe			10. Klassenstufe		
	argu.	info.	narr.	argu.	info.	narr.
Hauptschulen *n* = 262; 122	422 (68; 4.2)	458 (75; 4.7)	439 (64; 4.0)	461 (59; 5.3)	495 (75; 6.8)	483 (62; 5.7)
MBG *n* = 94; 69	464 (56; 5.8)	504 (66; 6.8)	474 (67; 7.0)	503 (63; 7.6)	524 (84; 10.1)	503 (55; 6.6)
IGS *n* = 203; 134	475 (73; 5.2)	479 (81; 5.7)	470 (68; 4.8)	496 (67; 5.8)	500 (69; 6.0)	496 (64; 5.5)
Realschulen *n* = 528; 345	506 (67; 2.9)	494 (68; 3.0)	509 (63; 2.7)	535 (60; 3.2)	511 (72; 3.9)	527 (64; 3.5)
Gymnasien *n* = 755; 484	553 (57; 2.1)	531 (67; 2.4)	541 (67; 2.4)	572 (62; 2.8)	545 (67; 3.0)	561 (63; 2.9)

Werte: Mittelwerte; Werte in Klammern: Standardabweichung und Standardfehler

4.3.2.2. *Stil*

Für die durchschnittliche Ausprägung von stilistischen Schreibfähigkeiten zeigen sich Gruppenunterschiede nach *Klassenstufe*, *Geschlecht*, *Sprachhintergrund* und *Schulform*. Die Unterschiede zeigen sich in allen drei Textmustern, teilweise jedoch in unterschiedlicher Stärke.

Der Unterschied hinsichtlich stilistischer Schreibkompetenzen zwischen den Klassenstufen 9 und 10 beträgt durchschnittlich 20–30 Punkte. Schülerinnen weisen im Schnitt um eine rund 30–40 Punkte erhöhte inhaltliche Schreibkompetenz gegenüber (männlichen) Schülern auf. Schülerinnen und Schüler mit deutscher Herkunftssprache zeigen einen Vorteil von 50–70 Punkten gegenüber Schülerinnen und Schülern mit anderen Herkunftssprachen (vgl. Tabelle 4.3.2.2.1). Es zeigen sich keine bedeutsamen textmusterspezifischen Unterschiede in der Stärke der Effekte (alle $Z < 1.96$).

Tabelle 4.3.2.2.1: ***Durchschnittliche stilistische Schreibkompetenzen nach Klassenstufe, Geschlecht und Sprachhintergrund.***

	Mittelwert argumentieren (SD; SE)	*p*-Wert Cohen's *d*	Mittelwert informieren (SD; SE)	*p*-Wert Cohen's *d*	Mittelwert narrativ (SD; SE)	*p*-Wert Cohen's *d*
9. Klassenstufe *n* = 1842	507 (78; 1.8)	< .001 0.41	506 (77; 1.8)	< .001 0.26	507 (77; 1.8)	< .001 0.35
10. Klassenstufe *n* = 1154	539 (72; 2.1)		525 (70; 2.1)		534 (73; 2.1)	
männlich *n* = 1471	501 (77; 2.0)	< .001 0.48	492 (74; 1.9)	< .001 0.58	502 (75; 2.0)	< .001 0.38
weiblich *n* = 1525	537 (71; 1.8)		534 (71; 1.8)		532 (75; 1.9)	
D als HS *n* = 2621	528 (73; 1.4)	< .001 0.86	522 (72; 1.4)	< .001 0.86	525 (74; 1.4)	< .001 0.88
¬ D als HS *n* = 374	461 (83; 4.3)		459 (72; 3.7)		461 (72; 3.7)	

SD: Standardabweichung; SE: Standardfehler des Mittelwerts

Im Vergleich der Schulformen zeigen sich keine textmusterspezifischen Unterschiede hinsichtlich der Effektmuster in Klassenstufe 9. Alle Schulformen erweisen sich als statistisch bedeutsam voneinander verschieden (alle $p < .05$) mit Ausnahme der MBG und der IGS, welche eine statistisch homogene Gruppe bilden. Die stilistische Schreibkompetenz der entsprechenden Schülerinnen und Schüler steigt in der Abfolge *Hauptschulen* – {*MBG*; *IGS*} – *Realschulen* – *Gymnasien* an (vgl. Tabelle 4.3.2.2.2). In Klassenstufe 10 erweisen sich mit einer Ausnahme dieselben Effekte als bedeutsam, lediglich der Unterschied zwischen den MBG und den Realschulen erreicht im informierenden Textmuster keine statistische Signifikanz ($p = .219$). Hinsichtlich der Stärke der Effekte erweisen sich die Unterschiede im argumentierenden Textmuster als stärker als diejenigen im informierenden Textmuster; das narrative Textmuster unterscheidet sich nicht bedeutsam von den beiden anderen.

Tabelle 4.3.2.2.2: ***Schulformbezogene durchschnittliche stilistische Schreibkompetenzen nach Klassenstufe.***

	9. Klassenstufe			10. Klassenstufe		
	argu.	info.	narr.	argu.	info.	narr.
Hauptschulen *n* = 262; 122	426 (67; 4.2)	437 (71; 4.4)	439 (63; 3.9)	468 (62; 5.6)	486 (73; 6.6)	476 (64; 5.8)
MBG *n* = 94; 69	467 (58; 5.9)	479 (66; 6.8)	466 (66; 6.8)	505 (64; 7.7)	501 (70; 8.4)	499 (58; 7.0)
IGS *n* = 203; 134	472 (77; 5.4)	469 (77; 5.4)	472 (67; 4.7)	498 (69; 5.9)	493 (71; 6.1)	500 (64; 5.5)
Realschulen *n* = 528; 345	507 (68; 3.0)	499 (64; 2.8)	498 (60; 2.6)	539 (62; 3.3)	515 (64; 3.5)	519 (65; 3.5)
Gymnasien *n* = 755; 484	551 (57; 2.1)	549 (62; 2.2)	552 (67; 2.4)	572 (62; 2.8)	557 (60; 2.7)	573 (63; 2.9)

Werte: Mittelwerte; Werte in Klammern: Standardabweichung und Standardfehler

4.3.2.3. *Sprachliche Richtigkeit*

Für die durchschnittliche Ausprägung von orthografisch-grammatischen Schreibfähigkeiten zeigen sich Gruppenunterschiede nach *Klassenstufe*, *Geschlecht*, *Sprachhintergrund* und *Schulform*. Die Unterschiede zeigen sich in allen drei Textmustern, teilweise jedoch in unterschiedlicher Stärke.

Der Unterschied hinsichtlich orthografisch-grammatischer Schreibkompetenzen zwischen den Klassenstufen 9 und 10 beträgt durchschnittlich 20–30 Punkte. Schülerinnen weisen im Schnitt um eine rund 40–50 Punkte erhöhte orthografisch-grammatische Schreibkompetenz gegenüber (männlichen) Schülern auf. Schülerinnen und Schüler mit deutscher Herkunftssprache zeigen einen Vorteil von 50–70 Punkten gegenüber Schülerinnen und Schülern mit anderen Herkunftssprachen (vgl. Tabelle 4.3.2.3.1). Es zeigen sich keine bedeutsamen textmusterspezifischen Unterschiede in der Stärke der Effekte (alle $Z < 1.96$).

Tabelle 4.3.2.3.1: ***Durchschnittliche orthografisch-grammatische Schreibkompetenzen nach Klassenstufe, Geschlecht und Sprachhintergrund.***

	Mittelwert argumentieren (SD; SE)	*p*-Wert Cohen's *d*	Mittelwert informieren (SD; SE)	*p*-Wert Cohen's *d*	Mittelwert narrativ (SD; SE)	*p*-Wert Cohen's *d*
9. Klassenstufe *n* = 1842	509 (81; 1.9)	< .001	510 (82; 1.9)	< .001	508 (77; 1.8)	< .001
		0.27		0.36		0.30
10. Klassenstufe *n* = 1154	531 (74; 2.2)		537 (71; 2.1)		531 (71; 2.1)	
männlich *n* = 1471	494 (78; 2.0)	< .001	493 (77; 2.0)	< .001	496 (74; 1.9)	< .001
		0.60		0.71		0.56
weiblich *n* = 1525	539 (73; 1.9)		545 (72; 1.8)		537 (72; 1.8)	
D als HS *n* = 2621	526 (75; 1.5)	< .001	529 (75; 1.5)	< .001	524 (74; 1.4)	< .001
		0.94		0.96		0.77
¬ D als HS *n* = 374	454 (78; 4.0)		456 (77; 4.0)		468 (72; 3.7)	

SD: Standardabweichung; SE: Standardfehler des Mittelwerts

Im Vergleich der Schulformen zeigen sich für die 9. Jahrgangsstufe keine textmusterspezifischen Unterschiede hinsichtlich der Effektmuster. Alle Schulformen erweisen sich als statistisch bedeutsam voneinander verschieden (alle $p < .05$) mit Ausnahme der MBG und der IGS, welche eine statistisch homogene Gruppe bilden. Die orthografisch-grammatische Schreibkompetenz der entsprechenden Schülerinnen und Schüler steigt in der Abfolge *Hauptschulen* – {*MBG*; *IGS*} – *Realschulen* – *Gymnasien* an. In der 10. Jahrgangsstufe findet sich dieses Muster nur für *argumentieren* wieder. Für *informieren* zeigt sich kein Unterschied mehr zwischen den Hauptschulen und den MBG sowie den IGS, für *narrativ* erweisen sich die Realschulen als nicht different von den MBG sowie den IGS (vgl. Tabelle 4.3.2.3.2). Hinsichtlich der textmusterspezifischen Effektstärkenvergleiche zeigt sich in Klassenstufe 9 kein bedeutsamer Unterschied, in Klassenstufe 10 erweist sich der Schulformeffekt für *argumentieren* als stärker als für die beiden anderen beiden Textmuster.

Tabelle 4.3.2.3.2: ***Schulformbezogene durchschnittliche orthografisch-grammatische Schreibkompetenzen nach Klassenstufe.***

	9. Klassenstufe			10. Klassenstufe		
	argu.	info.	narr.	argu.	info.	narr.
Hauptschulen *n* = 262; 122	420 (59; 3.6)	414 (61; 3.8)	424 (58; 3.6)	447 (54; 4.9)	476 (68; 6.2)	461 (59; 5.4)
MBG *n* = 94; 69	444 (59; 6.1)	454 (65; 6.7)	471 (62; 6.4)	471 (51; 6.1)	479 (57; 6.9)	501 (50; 6.0)
IGS *n* = 203; 134	453 (65; 4.6)	443 (68; 4.8)	470 (58; 4.0)	473 (63; 5.4)	476 (66; 5.7)	498 (58; 5.0)
Realschulen *n* = 528; 345	497 (58; 2.5)	505 (55; 2.4)	490 (55; 2.4)	517 (56; 3.0)	528 (55; 2.9)	507 (59; 3.2)
Gymnasien *n* = 755; 484	573 (52; 1.9)	571 (50; 1.8)	566 (57; 2.1)	585 (49; 2.2)	585 (49; 2.2)	579 (56; 2.6)

Werte: Mittelwerte; Werte in Klammern: Standardabweichung und Standardfehler

4.4. Zusammenfassung und Einordnung der Ergebnisse

Für alle drei Textmuster findet sich ein jahrgangsbezogener Unterschied, der jedoch für die Subgruppe der Schülerinnen und Schülern, die den MSA oder das Abitur anstreben, für informierende Texte in den statistisch nicht bedeutsamen Bereich abfällt. Für *informieren* lässt sich somit kein bedeutsamer Leistungszuwachs von Klassenstufe 9 zu Klassenstufe 10 für die relevante Schülergruppe verzeichnen. Dies steht im Einklang mit Ergebnissen aus der DESI-Studie, in welcher nur ein sehr geringer Leistungszuwachs (5–6 Punkte) für die semantisch-pragmatische Dimension und kein Leistungszuwachs für die sprachsystematische Dimension beim Vergleich von Schülerinnen und Schülern am Anfang und am Ende der Klassenstufe 9 gefunden wurde (Klieme et al. 2006; Klieme 2006; A. Neumann, 2007). Für die anderen beiden Textmuster lassen sich Zuwächse von 10 Punkten für die Narration und 22 Punkten für die Argumentation verzeichnen. Das argumentierende Muster hebt sich hierbei bedeutsam von den anderen Textmustern ab. Die Leistungen für alle drei Textmuster bewegen sich hinsichtlich ihres Zuwachses in einem Rahmen, welcher für andere sprachliche Kompetenzen zwischen den Klassenstufen 9 und 10 gefunden wurden. Diese reichen von keinen bis minimalen Fortschritten für *Lesen* und für die bereits genannten Bereiche *Sprachsystematik* und *semantisch-pragmatische Teilkompetenz* in DESI (Klieme et al., 2006;

A. Neumann, 2007) über Effekte im Bereich von rund 15 Punkten, etwa in *Orthografie* und *Lesen* in den IQB-Normierungsstudien (unveröffentlichte Daten), bis zu 30 Punkten für *Lesen* in PISA (Baumert, Stanat & Watermann, 2006; Ehmke, Klieme & Stanat, 2013; OECD, 2004). Unter normativer Interpretation und Klassifikation der Leistungswerte als Kompetenzstufen im Sinne von Mindeststandards, Regelstandards etc. zeigt sich neben dem stärkeren Leistungszuwachs für *argumentieren* auch eine bessere absolute Leistung für dieses Textmuster in den Klassenstufen 9 und 10. Vermutlich ist dies darauf zurückzuführen, dass Textsorten des Argumentierens[26] in den Klassenstufen 9 und 10 curricular zentral werden und sich die Schülerinnen und Schüler somit aktuell mit diesem Textmuster am meisten beschäftigen und innerhalb eines Jahres auch den größten Expertisezuwachs erwerben, während die anderen Textmuster, vor allem klassische Muster der Narration, keine oder nur (noch) periphere curriculare Relevanz (mehr) aufweisen. Eine curriculare Analyse von Harsch und Kollegen, die allerdings bereits 2001 durchgeführt wurde, bestätigt den Anstieg des Einsatzes argumentativer Textsorten ab Klassenstufe 8 (Harsch et al., 2009).

Die jahrgangsbezogenen Unterschiede spiegeln sich in allen drei Dimensionen *Inhalt*, *Stil* und *sprachliche Richtigkeit* in einer vergleichbaren Höhe von rund 20 bis 30 Punkten wieder.

Geschlechtereffekte erweisen sich mit rund 40 Punkten mit einem Vorteil der Schülerinnen als stärker als der Leistungszuwachs innerhalb eines Schuljahres. Die Höhe des Unterschieds entspricht dem mittleren Geschlechtsunterschied, der in Höhe von 41 Punkten in DESI für sprachliche Kompetenzen gefunden wurde. Im Vergleich zu anderen sprachlichen Kompetenzen, welche im Rahmen der Überprüfung des Erreichens der Bildungsstandards im Fach *Deutsch* getestet wurden, liegen die Effekte zwischen den rezeptiven Kompetenzen *Lesen* und *Zuhören*, wofür Geschlechtsdisparitäten in Höhe von knapp 20 Punkten nachgewiesen wurden, und dem teils produktiven, teils deklarativ-reflektorischen Kompetenzbereich *Orthografie* mit etwa 60 Punkten Geschlechterdifferenz (Winkelmann & Groeneveld, 2010). Die Geschlechterunterschiede finden sich in allen drei Schreibkompetenzdimensionen wieder, jedoch ansteigend von *Inhalt* (20–30 Punkte) über *Stil* (30–40 Punkte) bis zu *sprachliche Richtigkeit* (40–50 Punkten). Die Dimension *sprachliche Richtigkeit* weist somit ähnliche hohe Geschlechtereffekte auf wie die für den Kompetenzbereich *Orthografie* gefundenen.

[26] Der Ausdruck *Textsorte* bezieht sich hierbei auf spezifischere Textklassifikationsebenen als die Textmusterebene, vgl. hierzu auch Kapitel 6.2.; Textsorten des Argumentierens im schulischen Kontext sind beispielweise Bewerbungsschreiben oder Erörterungen.

Effekte des Sprachhintergrunds zeigen sich in Höhe von 50 bis 80 Punkten zugunsten der Schülerinnen und Schüler deutscher Herkunftssprache. Das argumentierende Textmuster weist hier den stärksten, das informierende den geringsten Effekt auf. Diese Effekte sind deutlich geringer als in DESI, wo ein Unterschied in den sprachlichen Leistungen in *Deutsch* (*Deutsch Gesamtleistung*) von 92 Punkten vorlag (Klieme et al., 2006). Im IQB-Ländervergleich zeigten vergleichende Analysen nach dem mit *Sprachhintergrund* assoziierten Merkmal *Migrationsstatus* Unterschiede zwischen 57 und 91 Punkten in den sprachlichen Kompetenzbereichen *Lesen*, *Zuhören* und *Orthografie* zwischen Jugendlichen ohne Migrationshintergrund und solchen, die in erster oder zweiter Generation in Deutschland wohnen (Böhme, Tiffin-Richards, Schipolowski & Leucht, 2010). Die hierbei gefundenen Unterschiede für *Orthografie* in Höhe von 50–70 Punkten spiegeln sich in gleicher Höhe auf der Schreibkompetenzdimension *sprachliche Richtigkeit* wider. Für *Stil* liegen Effekte in vergleichbarer Höhe vor, die Effekte für *Inhalt* sind mit 40 bis 60 Punkten etwas schwächer.

Schulformeffekte zeigen sich für die Schreibkompetenzen in allen drei Textmustern und auch für alle drei Kompetenzdimensionen. Hierbei steigen die Kompetenzen von Hauptschulen über Realschulen bis Gymnasien durchweg bedeutsam an. Schulen mit mehreren Bildungsgängen und Integrierte Gesamtschulen gliedern sich hierbei, teilweise bedeutsam in die eine und/oder andere Richtung verschieden, zwischen die Haupt- und Realschulen ein. Mit einem Maximalunterschied zwischen den Randkategorien *Hauptschulen* und *Gymnasien* von bis zu 170 Punkten zeigt sich ein ähnliches Bild wie für sprachliche Leistungen in *Deutsch* in anderen Schulleistungsstudien. In DESI wurden Schulformeffekte für die Gesamtleistung in *Deutsch* von knapp über 180 Punkten (Klieme et al. 2006), in PISA 2003 für die Lesekompetenz von rund 190 Punkten im Vergleich von Hauptschulen und Gymnasien gefunden (PISA-Konsortium, 2004). Textmusterspezifische Unterschiede in der Stärke der Effekte zeigen sich auch hier mit stärkeren Differenzen im Argumentieren (teilweise als in der Narration) als im Informieren. Auch diesbezüglich wird vermutet, dass diese Effekte curriculare Ursachen haben. So zeigt die curriculare Analyse von Harsch und Kollegen, dass argumentative Textsorten am Ende der Sekundarstufe I in Gymnasien stärker eingesetzt werden als in Haupt- und Realschulen. Dies trifft teilweise auch auf die Narration zu, vor allem auf elaboriertere narrative Formen wie das kreative Schreiben.

Auffallend ist, dass auch für die Dimension *sprachliche Richtigkeit*, die über alle Textmuster hinweg auf derselben Skala bewertet wurde, textmusterspezifische Unterschiede in den Effektstärken auftreten (i. e. stärkere schulformbezogene Unterschiede in Klassenstufe 10 für

argumentieren im Vergleich zu den anderen beiden Textmustern). Auch dies lässt sich vermutlich auf Ursachen der curricularen Verankerung zurückführen. Ist die Aufmerksamkeit des Schreibenden stärker durch den Versuch der inhaltlichen und stilistischen Bewältigung der Aufgabe gebunden, stehen weniger kognitive Kontrollinstanzen zum orthografischen und grammatischen Prüfen des Geschriebenen zur Verfügung (Heim, Keil & Ihssen, 2006; Ruland, Willmes & Günther, 2012). Aufgrund der stärkeren Routine und höheren Expertise im Argumentieren in den höheren Schulformen (Harsch et al., 2009), könnte dies eine mögliche Ursache für den stärkeren Unterschied im Rahmen der Argumentation darstellen.

5. Validität

In diesem Kapitel wird der messtheoretische Rahmen für die nachfolgenden drei Forschungskapitel dieser Arbeit dargelegt. Alle drei Teilstudien, welche in den Kapiteln 6 bis 8 vorgestellt werden, beschäftigen sich mit Validitätsaspekten bei der Messung von Schreibkompetenzen. Das vorliegende Kapitel ist wie folgt gegliedert: Kapitel 5.1. beleuchtet zunächst das Konzept von Validität. Hierbei wird ein kurzer Überblick über unterschiedliche Positionen und deren Verhältnis zueinander gegeben. Kapitel 5.2. geht auf verschiedene Formen der Validität bzw. auf Validitätsaspekte ein. In Kapitel 5.3. erfolgt schließlich ein Ausblick auf die Fragestellungen der Kapitel 6 bis 8 und setzt diese zu den betreffenden Validitätsaspekten in Beziehung.

5.1. Das Konzept *Validität*

Validität ist neben der Objektivität (Unabhängigkeit der Messung / der Testergebnisse vom Experimentator) und Reliabilität (Zuverlässigkeit der Messung / der Ergebnisse) eines der drei Hauptgütekriterien empirischer Forschung (Bortz & Döring, 2006; Bühner, 2014; Fisseni, 2004; Schmidt-Atzert & Amelang, 2012; Wilhelm & Kunina, 2009; Wolfer, 2010)

Die Validität gibt hierbei das Ausmaß an, „in dem ein Test misst, was er zu messen vorgibt.“ (Bühner, 2004, S. 30, zurückgehend auf Ruch, 1924; vgl. auch Bortz & Döring, 2006; Hartig, Frey & Jude, 2008; Wilhelm & Kunina, 2009; Wolfer, 2010). Nach Murphy und Davidshofer (2001) trifft diese Definition streng genommen nur auf einen Validitätsaspekt in der klassischen Untergliederung in Inhalts-, Kriteriums und Konstruktvalidität zu, die Inhaltsvalidität (siehe Kapitel 5.2.). Bühner (2004) weist darauf hin, dass die anderen beiden Validitätsaspekte (siehe Kapitel 5.2.) keine Eigenschaft des Tests, sondern eine Eigenschaft von auf den Testergebnissen basierenden Interpretationen ist.

De facto besteht keine Einigkeit darüber, auf welcher Ebene der Validitätsbegriff als solcher anzusiedeln ist, sprich, ob sich *Validität* auf Tests, auf Testergebnisse (Testwerte), auf Interpretationen von Testergebnissen oder gar auf weitere Aspekte bezieht (Newton & Shaw, 2012, 2014). So sprechen sich Messick (1989, 1990, 1995b), Chronbach (1988, 1989) und Kane (1992, 2002, 2006, 2013a, 2013b) für einen funktionalen Validitätsbegriff aus, der sich auf die Interpretation und Verwendung von Testergebnissen bezieht. Nach Messick (1995a;

1995b) ist Validität „die allgemeine Bewertung des Grades, in dem empirische und theoretische Evidenz für die Angemessenheit von Interpretationen und Handlungen, die auf diagnostischen Messungen beruhen, vorliegt“ (Gärtner & Pant, 2011, S. 11). Auch die Standards der American Psychological Association (APA, 2002) lehnen sich an Messick an, interpretieren Validität jedoch als Testwerteigenschaft.

> „In den APA-Standards (2002) wird Validität als Eigenschaft der Testwerte verstanden. Validität gibt den Grad an, zu dem die empirischen Belege und theoretischen Sachverhalte die beabsichtigten Interpretationen der Testwerte unterstützen.“ (Wilhelm & Kunina, 2009, S. 314)

Für Borsboom, Mellenbergh und van Heerden (2004) ist ein Konzept, wie es Messick oder Kane vertreten, welches sich auf Interpretationen und Handlungen bezieht und selbst ethische Aspekte miteinbezieht (Messick, 1980, 1981, 1989, 1990, 1995a; Kane 2013a, 2013b), zu weit. Nach Bornsboom und Kollegen ist Validität eine Eigenschaft eines Tests; der Test misst eine bestimmte Eigenschaft und ist dann valide, wenn eine Änderung der Eigenschaft mit einer Änderung im Testergebnis einhergeht. Borsboom und Kollegen gehen in ihrer Konzeption explizit von beobachtungsunabhängigen Eigenschaften sowie einem kausalen Zusammenhang zwischen diesen ‚wahren’ Eigenschaften und den Testergebnissen aus (Borsboom, Cramer, Kievit, Zand Scholten & Franic, 2009; Borsboom & Markus, 2013).

Hood, der sich selbst ebenfalls für einen wissenschaftstheoretischen Realismus ausspricht, führt aus, dass die Theorien von Messick einerseits und Borsboom et al. anderseits keinen Widerspruch darstellen müssen, er sieht die Theorien vielmehr als komplementär:

> „Borsboom et al. contribute semantic and ontological components while Messick provides the methodological tools for constructing an epistemology of psychological measurement. “ (Hood, 2009, S. 451)

Während Borsboom und Kollegen also eine logisch-semantische und ontologische Fundierung liefern, was Validität ist bzw. wie sie konzeptualisiert werden sollte, bietet Messicks Theorie Hinweise zu Verfahren, wie Validität überprüft werden kann.

Yousfi (2001) weist jedoch darauf hin, dass man die Borsboomsche Validitätskonzeption auch annehmen kann, ohne eine wissenschaftstheoretisch realistische Position zu vertreten:

> „Auch wenn Borsboom et al. ausdrücklich eine realistische erkenntnistheoretische Position beziehen, ist der von ihnen vertretene Validitätsbegriff durchaus auch mit einer konstruktivistischen oder idealistischen Sichtweise vereinbar. Die Überlegungen von Borsboom et al. machen auch dann Sinn, wenn man das Postulat von der Existenz des latenten Merkmals aufgibt und lediglich fordert, dass die Theorien, die wir zur Erklärung empirischer Phänomene heranziehen, latente Variablen enthalten, die als kausale Ursachen

> eben dieser Phänomene aufgefasst werden. Die latenten Variablen müssen also nicht unbedingt tatsächlich existieren, sondern es reicht, wenn sie innerhalb psychologischer Theorien als kausale Ursache für die Messwerte betrachtet werden.“ (S. 162)

Eine Vereinbarkeit der Aspekte der verschiedenen Validitätskonzeptionen nehmen auch Newton und Shaw (2014) implizit an, indem sie ein Framework wissenschaftlicher *Evaluation* vorschlagen, welches die verschiedenen Aspekte der vermeintlich konkurrierenden Konzeptionen integriert, dabei allerdings explizit auf den Begriff *Validität* verzichtet. Den Autoren zufolge ist die Streitfrage, was unter *Validität* zu verstehen sei, eine rein begriffliche, deren Beantwortung empirisch nicht entscheidbar sei, sondern sich letztendlich durch den konkreten Gebrauch des Begriffes innerhalb der wissenschaftlichen Gemeinschaft etabliere.

5.2. Validitätsaspekte

Klassischerweise werden drei (Haupt-)Validitätsaspekte unterschieden: Inhalts-, Kriteriums- und Konstruktvalidität (Blömeke, 2013; Bortz & Döring, 2006; Bryant, 2000; Bühner, 2004; Hartig et al., 2008; Kubinger, 2009).

Unter *Inhaltsvalidität* versteht man die Genauigkeit, mit welcher ein Test die Inhalte eines betreffenden Konstrukts erfasst. Inhaltsvalidität wird in der Regel durch Expertentum festgelegt und erfolgt anhand logischer und fachlicher Überlegungen. (Blömeke, 2013; Bühner, 2004; Hartig et al., 2008). Hartig und Kollegen unterschieden hierbei zwischen Inhaltsvalidität bei operational definierten Konstrukten und Inhaltsvalidität bei theoretischen Konstrukten. Bei theoretisch definierten Konstrukten basieren die logisch-fachlichen Überlegungen auf Einschätzungen zur Übereinstimmung von theoretischem Konstrukt und der vom Test erfassten Fähigkeiten und Eigenschaften; deshalb spricht man hier auch von *logischer Validität* (Bühner, 2004). Wenn auch Laien und/oder die Testpersonen inhaltlich anhand des Tests erkennen können, was mit diesem gemessen wird, spricht man auch von *Augenscheinvalidität* oder *psychologischer Validität* (Bühner, 2004; Fisseni, 2004). Bei operational definierten Konstrukten, sprich in Fällen, in denen der Test bzw. die Itemmenge das Konstrukt definiert, erübrigt sich ein Vergleich von Eigenschaften des theoretischen Konstrukts und durch den Test erfassten Merkmalen. Hier betreffen Überlegungen zur inhaltlichen Validität den Aspekt, ob die mit dieser Itemmenge generierten Testwerte und deren Interpretation verallgemeinerbar sind und die gewählten Items die Menge der

möglichen Items hinreichend repräsentieren. Häufig wird für diesen Aspekt auch der Ausdruck *repräsentative Validität* benutzt (Wolfer, 2010). Entsprechen die durch den Test erfassten Merkmale vollständig dem interessierenden Konstrukt wird auch von *trivialer Validität* gesprochen (Kubinger, 2009).

Bei der *Kriteriumsvalidität* handelt es sich um den Zusammenhang zwischen dem in einem Test gemessenen Kriterium und einem testexternen Kriterium. Bei Letzterem kann es sich bspw. um ein Beobachtungsdatum oder auch um das Ergebnis eines bereits etablierten anderen Tests, der dasselbe Konstrukt misst, handeln (Bortz & Döring, 2006; Blömeke, 2013; Bühner, 2004; Hartig et al., 2008; Schnell, Hill & Esser 2011). In diesem Rahmen werden hauptsächlich zwei Formen unterschieden, die *konkurrente* oder *Übereinstimmungsvalidität*, bei welcher die beiden zu vergleichenden Kriterien annähernd zeitgleich erhoben wurden und die *prognostische*, *prädikative* oder *Vorhersagevalidität*, bei welcher die Erhebung des Testkriteriums der Erhebung des Vergleichskriteriums zeitlich vorausgeht. Bei hinreichender Übereinstimmung hat das Testkriterium hinreichend prädikative oder prognostische Validität. Für den Fall, dass das Vergleichskriterium zeitlich vor dem Testkriterium erhoben wurde, spricht man auch von *retrospektiver Validität* (Bühner, 2004). Darüber hinaus kann einem Kriterium *inkrementelle Validität* zugeschrieben werden. Darunter versteht man den zusätzlichen Anteil, den das Testkriterium im interessierenden Konstrukt über bereits bekannte Maße und Kriterien hinaus erklärt (Bühner, 2004; Blömeke, 2013; Hartig et al., 2008).

Die *Konstruktvalidität* bezieht sich darauf, dass „mit Tests erfasste Merkmale immer konstruierte Größen sind, die auf hypothetische Konstrukte bezogen sein sollten.“ (Blömeke, 2013, S. 6). Sie bestimmt, in welchem Maße ein Test ein relevantes Merkmal in Einklang mit bestehenden theoretischen Annahmen über das Konstrukt misst. Hierbei wird vor allem die Nähe und Ferne des Konstrukts zu anderen Konstrukten berücksichtigt. Nach Cronbach und Mehl (1955) bilden sich in einem nomologischen Netz die theoretisch angenommenen Konstruktbeziehungen, welche als Axiome bezeichnet werden, über Korrespondenzregeln verknüpft als empirische Gesetze auf der Beobachtungsebene ab (Hartig et al., 2008). Campell und Fiske (1959) treffen hierbei eine bis heute relevante Differenzierung in *konvergente* und *divergente* oder *diskriminante Validität*. Unter konvergenter Validität wird verstanden, dass Messverfahren, welche dasselbe Konstrukt oder Konstrukte, die theoretisch eng miteinander verwandt sind, messen, hohe Zusammenhänge miteinander aufweisen sollten. Unter diskriminanter (oder divergenter) Validität wird verstanden, dass

Messverfahren, welche unterschiedliche und nicht als eng verwandt angenommene Konstrukte abbilden, niedrige Zusammenhänge ausweisen sollten.

Neben den externen Beziehungen eines Konstrukts zu anderen Konstrukten ist auch die interne Struktur des Konstrukts für die Konstruktvalidität relevant. So sollte sich die theoretisch angenommene dimensionale Struktur in den durch den Test gewonnenen Daten widerspiegeln. Dieser Validitätsaspekt wird als *faktorielle Validität* (Hartig et al., 2008) oder *strukturelle Validität* (Gärtner & Pant, 2011; Hadenfeldt & Neumann, 2012; Roick, 2008) bezeichnet.

Während zahlreiche Autoren sich bis heute an der Dreiteilung von *Inhalts-*, *Kriteriums-* und *Konstruktvalidität* orientieren (Blömeke, 2013; Bortz & Döring, 2006; Bühner, 2004; Hartig et al., 2008; Kubinger, 2009; Schnell et al., 2011), plädiert Messick (1989, 1990, 1995b; auch Anastasi, 1986) für ein einheitliches Validitätskonzept, nämlich ein erweitertes Verständnis von Konstruktvalidität (vgl. Kapitel 5.1. zu Messicks Validitätskonzept). Aspekte der inhaltlichen und kriterialen Validität lassen sich unter diese subsumieren; einige Autoren verzichten daher in Anlehnung an Messick auf diese Unterteilung (Wilhelm & Kunina, 2009; APA, 2002). Kane (2013a) stellt heraus, dass Messicks einheitliche Theorie der Konstruktvalidität eine theoretisch elegante Lösung war, allerdings wenig Praktisches und Methodisches zur Überprüfung von Konstruktvalidität bereitstellte. Auch wenn Messick einerseits auf oberster theoretischer Ebene eine einheitliche (Konstrukt-)Validitätstheorie vertritt, führte er andererseits eine detailliertere Untergliederung in Validitätsaspekte als die klassische Dreiteilung ein (Messick, 1990; 1995b). So unterscheidet er in *Inhaltsvalidität*, *substantielle Validität*, *strukturelle Validität*, *Generalisierbarkeit*, *externale Validität* und *Konsequenz-* oder *Folgevalidität*.[27] *Inhaltsvalidität* und *strukturelle Validität* entsprechen den bereits erläuterten Konzepten. *Externale Validität* bezieht sich auf die Zusammenhänge mit testexternen Kriterien, sie beinhaltet die klassische Kriteriumsvalidität sowie die konstruktübergreifenden Teilaspekte der Konstruktvalididtät, i. e. diskriminante und konvergente Validität. Unter *substantieller Validität* wird der Anspruch verstanden, dass nicht nur die Testergebnisse bzw. deren Interpretation mit theoretischen Annahmen übereinstimmen müssen, sondern dass diese Ergebnisse auch über identische (kognitive) Prozesse zustande kommen, wofür selbst empirische Evidenzen nötig sind. Unter *Generalisierbarkeit* versteht Messick das Ausmaß, in welchem sich Testwerteigenschaften und -interpretationen

[27] Messick selbst spricht aufgrund seiner einheitlichen Validitätstheorie hier jeweils nur von Validitätsaspekten und nicht von spezifischen Validitäten (Messick, 1990, 1994). Die Begrifflichkeit von spezifischen Validitäten etablierte sich im Rahmen der Messickrezeption und Fortführung seiner Arbeiten (u. a. Gärtner & Pant, 2011; Wasserman & Bracken, 2003; Reckase, 1998; Lees-Haley, 1996).

über Gruppen, Situationen und Aufgaben generalisieren lassen. Unter *Konsequenz-* oder *Folgevalidität* thematisiert Messick die Kurz- und Langzeitwirkungen der Testwert-interpretationen sowie der Testanwendung. Wie bereits in Kapitel 5.1. erläutert, spielen hier auch ethische und soziale Faktoren eine Rolle, weshalb dieser Punkt als Validitätsaspekt nicht unumstritten ist.

Unabhängig von der Zustimmung zu Messicks einheitlicher Theorie von (Konstrukt-) Validität oder der Annahme der Differenzierung bestimmter Validitätsaspekte hat sich eine von Messick ausgearbeitete Unterscheidung von Hauptgefahren bzw. -gefährdungen von Konstruktvalidität etabliert. So kann aufgrund der *Unterrepräsentation des Konstrukts* im Rahmen einer Testung die Messung zu eng sein, indem nicht alle Aspekte des Konstrukts hinreichend erfasst werden. Das Gegenstück hierzu ist die *konstruktirrelevante Varianz*, worunter die Erfassung konstruktexterner Aspekte zu verstehen ist; die Messung ist unter Vorliegen von konstruktirrelevanter Varianz zu breit (Messick, 1990; 1995b; auch: Bühner, 2004; Caspari, Grotjahn & Kleppin, 2010; Yousfi, 2011).

Eine weitere Unterscheidung von Validitätsaspekten, die sich allerdings auf gesamte Versuchspläne und das Verhältnis von abhängigen und unabhängigen Variablen und die Möglichkeit kausaler Interpretationen bezieht, ist die zwischen *interner* und *externer Validität*. Von *interner Validität* wird gesprochen, wenn bei der Experimentaldurchführung Störeinflüsse und somit Alternativerklärungen (zur Erklärung auf Basis der Variation in der unabhängigen Variablen) für die Variation der abhängigen Variablen ausgeschlossen werden kann. *Externe Validität* bezieht sich auf die Repräsentativität und Generalisierbarkeit der Ergebnisse einer Untersuchung. Hier wird auch zwischen *Populationsvalidität*, welche sich darauf bezieht, dass die Stichprobenergebnisse auf andere Personen, Personengruppen oder Populationen übertragen bzw. verallgemeinert werden können, und *Situationsvalidität* oder *ökologische Validität* unterschieden. Letztere bezieht sich darauf, inwieweit die Ergebnisse auf andere Kontexte übertragen bzw. situationsübergreifend verallgemeinert werden können. (Bortz & Döring, 2006; Deinzer, 2007; Sarris & Reiß, 2005; Schnell et al., 2011; Seel, Pirnay-Dummer & Ifenthaler, 2010).

5.3. Validitätsaspekte in den Folgeuntersuchungen

Die folgenden drei Kapitel 6 bis 8 widmen sich spezifischen Hauptfragestellungen, die sich alle als Validitätsfragen im Rahmen der Schreibkompetenzerfassung auffassen lassen.

In Kapitel 6 wird der Frage nachgegangen, ob man basierend auf den empirischen Daten von einem textmusterunabhängigen Konstrukt *Schreibkompetenz* ausgehen kann oder ob sich Evidenzen für textmusterspezifische Schreibkompetenzen finden. Oder anders ausgedrückt: Handelt es sich bei *Schreibkompetenz* um ein textmusterübergreifendes und in dieser Hinsicht eindimensionales Konstrukt oder um ein mehrdimensionales Konstrukt mit textmusterspezifischen Schreibkompetenzdimensionen? Die Fragestellung betrifft somit die interne Struktur des Konstrukts und damit die faktorielle oder strukturelle Validität (und somit sowohl im klassischen als auch im Messickschen Sinne die Konstruktvalidität).

Leitfrage in Kapitel 7 ist, in welchem Umfang Lesekompetenz bei der Erfassung der Schreibkompetenz aufgrund der schriftlichen Schreibaufgabenpräsentation (unintendierterweise) mitgemessen wird. Auch diese Frage betrifft wiederum die Konstruktvalidität; hierbei geht es im Sinne Messicks um die Ermittlung des Anteils an (spezifischer) konstruktirrelevanter Varianz. Im Rahmen dieser Untersuchung werden die Schreibleistungswerte in Abhängigkeit von leseschwierigkeitsbestimmenden Merkmalen der entsprechenden Schreibaufgabenstimuli mit Lesekompetenzwerten verglichen, weshalb hier Aspekte der diskriminanten Validität bzw. der externale Validitätsaspekt nach Messick zum Tragen kommen.

In Kapitel 8 wird ebenfalls das Problem der konstruktirrelevanten Varianz und somit ein Aspekt der Konstruktvalidität beleuchtet, hier allerdings auf der Ebene von konzeptuell unabhängigen Schreibkompetenzdimensionen, i. e. inhaltlicher, stilistischer und orthografisch-grammatischer Schreibkompetenzen. Hier werden mögliche Halo-Effekte bei der Beurteilung von inhaltlichen und stilistischen Schreibkompetenzen in Abhängigkeit von der sprachlichen Richtigkeit dieser Texte untersucht. Hierbei wurde die sprachliche Richtigkeit als vermeintliche Halo-Effekte induzierende Quelle für die Vergleichsbedingungen experimentell variiert, sodass Unterschiede zwischen den Bedingungen auf diese konstruktirrelevante Größe zurückgeführt werden können.

6. Textmusterspezifität vs. Textmusterunabhängigkeit von Schreibkompetenzen (Teilstudie I)

Dieses Kapitel verfolgt die Fragestellung, ob es sich bei den Schreibkompetenzen von Schülerinnen und Schülern am Ende der Sekundarstufe I um textmusterspezifische oder textmusterunabhängige Fähigkeiten handelt. Dabei werden neben den (ganzheitlichen) Schreibkompetenzen auch die angenommen Teilkompetenzen, d. h. inhaltliche, stilistische und orthografisch-grammatische Schreibkompetenzen betrachtet sowie deren Verhältnis zueinander, i. e. die interne Struktur von *Schreibkompetenz*.

6.1. Textsortenkompetenz / Textmusterkompetenz

Verschiedene Textmuster wie das Informieren, das Argumentieren oder das Erzählen sind mit unterschiedlichen Herausforderungen an den Schreiber entsprechender Texte verbunden (Berman & Nir-Sagiv, 2007; Devitt, 2008; Knapp & Watkins, 2005; Risel, 2011; Ulmi, Bürki, Verhein & Marti, 2014).[28] Aus diesem Grund ist es eine weit verbreitete fachdidaktische Annahme, dass Schreibkompetenzen textsortenspezifisch sind. Empirische Belege hierfür finden sich – zumindest im deutschsprachigen Raum – jedoch nur wenige. So schreibt Feilke (2006, S. 183):

> „Es gibt eine überschaubare Zahl textsortendifferenzierender Untersuchungen (…). Untersuchungen zur Interdependenz von Textsorten in der Entwicklung gibt es bis heute nicht."

Selbst in diesen „überschaubare[n] textsortendifferenzierende[n] Untersuchungen" wird eine umfassende Textsortenkompetenz jedoch nicht auf eine Art und Weise untersucht, sodass diese Behauptung als hinreichend gestützt angesehen werden könnte. Ein Nachweis von textsortenspezifischen Kompetenzen sollte auf zweierlei Evidenzen beruhen:

(i) Bei der Bearbeitung mehrerer Aufgaben derselben Textsorte erweisen sich Schreibleistungen als stabil.
(ii) Bei der Bearbeitung mehrerer Aufgaben unterschiedlicher Textsorte zeigen sich differenziertere Leistungen.

[28] Weitere Ausführungen hierzu im Folgenden unter Kapitel 6.3.

Die meisten Untersuchungen im Bereich der Textsortenkompetenz konzentrieren sich auf ein Textmuster bzw. eine Textsorte und fokussieren dabei oftmals spezifische sprachliche Aspekte, die aufgrund theoretischer Überlegungen an eine Textsorte gebunden werden. So untersucht beispielweise Bachmann (2002) Kohäsion und Kohärenz an instruierenden Texten, Hug (2001) Temporalität an erzählenden Texten. Während diese Studien als Einzelevidenzen und Detailhinweise verstanden werden können, verweist Pohl (Augst, Disselhoff, Henrich, Pohl & Völzing, 2007, S. 29) jedoch auf ein wesentliches Hauptproblem dieser Studien, nämlich dass sie sich weitestgehend eben nur auf eine Textsorte stützen. Somit können textsortenbezogene Fähigkeiten, insofern sie denn in hinreichendem Maße nachgewiesen werden, nicht von generellen Schreibfähigkeiten unterschieden werden.

Augst et al. (2007) verfolgen einen textsortenkontrastierenden Ansatz und ziehen Schülerleistungsdaten von fünf Textsorten bzw. Textmustern heran: *erzählen*, *berichten*, *beschreiben*, *instruieren* und *argumentieren*, wobei *berichten*, *beschreiben* und *instruieren* Subtypen des Typs *informieren* bzw. *darstellen* sind. Die Studie von Augst und Kollegen hat jedoch zwei wesentliche Nachteile: Zum einen wird der Hauptbezugsrahmen, welcher der Kategorisierung der Texte (als Ganzes) zugrunde liegt, i. e. ontogenetische Stufenmodelle, textsortenspezifisch gewählt, d. h. ein gemeinsamer (textsortenübergreifender) Bezugsrahmen zum Vergleich der textsortenspezifischen Fähigkeiten fehlt. Zum anderen wird pro Textsorte jeweils nur eine Aufgabe herangezogen, d. h. eine Abgrenzung aufgabenspezifischer Effekte von textsortenspezifischen Effekten ist in diesem Rahmen nicht möglich, Schlussfolgerungen bezüglich Textsortenspezifität sind nicht hinreichend gestützt.

Auch wenn eine Trennung zwischen Textsorten- und Aufgabenspezifität in den Untersuchungen von Augst und Kollegen nicht möglich ist und somit obigem Punkt (i) nicht entsprochen werden kann, liefern die Autoren dennoch mit Blick auf Punkt (ii) einige Evidenzen für tendenzielle textsortenübergreifende Aspekte, beispielsweise im Hinblick auf die Verwendung rahmender Elemente oder die Textlänge, bezüglich anderer Aspekte (z. B. die Verwendung konditionaler Strukturen) textsortenspezifische Evidenzen.

Ein Blick in die Forschung des angelsächsischen Sprachraums zeigt, dass für das Englische ein breiteres Spektrum an Untersuchungen hinsichtlich der Textsortenspezifität vs. Textsortenunabhängigkeit von Schreibkompetenzen vorliegt (Carlman, 1985; Engelhard, Gordon & Gabrielson, 1992; Engelhard, Gordon, Gabrielson & Walker, 1994; Kegley, 1986; Olinghouse, Santangelo & Wilson, 2012; Prater & Padia, 1983; Quellmalz, Capell & Chou,1982; Shermis, Shneyderman & Attali, 2008; Veal & Tillmann, 1971). Diese Studien

liefern teilweise ebenfalls Evidenz für Unterschiede in den Leistungen der Schreibenden in Abhängigkeit des Textmusters bzw. der Textsorte des zu schreibenden Textes. Allerdings wurde auch im Rahmen der meisten der aufgeführten Studien wiederum nur eine Aufgabe pro Textmuster eingesetzt, weshalb eine Differenzierung zwischen Aufgabenunterschieden und Textmusterunterschieden in der Interpretation der Effekte nicht möglich ist. Jedoch liegen auch Untersuchungen vor, in welchen versucht wurde, diesen Effekt zu kontrollieren. So verglichen Quellmalz et al. (1982) die Korrelationen von aufgabenspezifischen Schreibleistungswerten von US-amerikanischen Schülerinnen und Schülern der elften und zwölften Jahrgangsstufe bei der Bearbeitung zweier textmusteridenter oder textmusterdifferenter Aufgaben, die verwendeten Aufgaben waren den Textmustern *erklären* und *erzählen* zuzuordnen. Dabei zeigte sich ein bedeutsam höherer Zusammenhang für die textmusteridentische Aufgabenbearbeitung als für die textmusterdifferente Aufgabenbearbeitung. Olinghouse et al. (2012) verfolgten einen anderen Ansatz und versuchten, die Unterscheidung zwischen Aufgabenspezifität und Textmusterspezifität aufzuheben, indem sie zwar nur eine Aufgabe pro Textmuster – verwendete Textmuster: *argumentieren*, *informieren*, *erzählen* – einsetzten, allerdings alle drei Aufgaben zum gleichen Themenfeld, „Weltraum“ (im Original: *outer space)*, gestellt wurden. Die Studie wurde an US-amerikanischen Schülerinnen und Schülern der fünften Jahrgangsstufe durchgeführt; jede Schülerin und jeder Schüler bearbeitete eine Aufgabe pro Textmuster (insgesamt drei Aufgaben). Hinsichtlich der erfassten holistischen Qualitätsurteile zeigten sich Korrelationen zwischen den aufgaben- und somit auch textmusterspezifischen Schreibleistungswerten zwischen .37 und .48 (S. 71). Die Autoren interpretieren diese Befunde als Evidenz für textmusterspezifische Schreibkompetenzen. Fraglich ist dabei jedoch, inwiefern der Versuch der Kontrolle der Aufgabenspezifität gelungen ist; bei genauerer Betrachtung unterscheiden sich die drei Aufgaben, auch wenn sie dem gleichen Großthema zuzuordnen sind, in den thematischen Anforderungen und den erforderlichen Wissensressourcen nicht unerheblich.[29] Darüber hinaus variieren die Aufgaben hinsichtlich anderer stilistischer Anforderungen, die unterhalb der Ebene des Textmusters oder auf anderer stilistischer Ebene anzusiedeln sind, so beispielsweise dem intendierten Adressaten und damit verbunden etwa dem Formalitätsgrad.

[29] So musste bei der Bearbeitung der argumentativen Aufgabe zu einem Vorhaben des US-amerikanischen Präsidenten, extraterrestrische Wohnräume anzulegen, Stellung bezogen werden. Im Rahmen der erzählenden Aufgabe sollte eine Geschichte verfasst werden unter dem Hintergrund, dass Astronauten etwas (= unbestimmt) unter der Mondoberfläche entdecken. Die informierende Aufgabe bestand aus dem Auftrag, die wichtigsten Aspekte über den Weltraum, welche der Schreiber selbst festzulegen hatte, jemandem (= unbestimmt) zu berichten.

Zusammenfassend lässt sich festhalten, dass bisher nur wenig systematische Evidenz für oder gegen textmusterspezifische (vs. textmusterunabhängige) Schreibkompetenz erbracht wurde. Thematische und stilistische aufgabenspezifische Effekte wurden in den meisten Studien nicht kontrolliert, die Menge der eingesetzten Aufgaben war meistens sehr gering. Dennoch lässt sich ein tendenzielles Muster ausmachen, welches für eine Textmusterspezifität von Schreibkompetenzen spricht.

6.2. Exkurs: *Textmuster*, *Textsorte* und Co. – Begriffe und Klassifikationen

In den Ausführungen unter 6.1. wurden die Begriffe *Textsorte* und *Textmuster* weitgehend synonym verwendet, auch um hinsichtlich der Erläuterungen zu Theorien und Befunde deutschsprachiger Autoren konsistent zu deren verwendeten Begrifflichkeiten zu bleiben. Ein Blick in die Literatur zeigt jedoch, dass es keine einheitliche Verwendungsweise dieser Begriffe gibt. Daneben besteht auch kein Konsens darüber, auf welcher Ebene der Textsortenbegriff anzusiedeln ist. Vielmehr erweist sich der Begriff als Operationsbegriff. So

> „arbeiten Textlinguisten sehr gern mit einem »alltagssprachlichen Textsortenbegriff«, da das »alltagssprachliche« Vokabular zur Bezeichnung von Textsorten bei einer für die kommunikative Praxis auch nur einigermaßen relevanten wissenschaftlichen Begriffsbildung unbedingt berücksichtigt werden müsse." (Gansel & Jürgens, 2009, S. 53).

Isenberg (1983) stellt verschiedene Möglichkeiten vor, wie man zu einer angemessen Text- und Textsortentheorie gelangen kann, unter anderem auch solche, den klassischen („alltagssprachlichen") Textsortenbegriff durch einen aus einer Texttheorie abgeleiteten definierbaren Begriff zu ersetzen. Diese Möglichkeiten wurden jedoch aus den von Gansel und Jürgens zusammengefassten Gründen (s.o.) nicht akzeptiert (vgl. auch Adamzik, 1991; Brinker, 1985).

Brinker (1985) lieferte eine etwas breiter akzeptierte Definition (u. a. Adamzik, 2004; Fischer, 2009; Gantefort, 2013; Lothar Hoffmann, 1998a, 1998b), die jedoch begrifflich sehr abstrakt und in weiten Teilen vage gehalten ist:

> „Textsorten sind konventionell geltende Muster für komplexe sprachliche Handlungen und lassen sich als jeweils typische Verbindungen von kontextuellen (situativen), kommunikativ-funktionalen und strukturellen (grammatischen und thematischen) Merkmalen beschreiben. Sie haben sich in der Sprachgemeinschaft historisch entwickelt und gehören zum Alltagswissen der Sprachteilhaber: sie besitzen zwar normierende Wirkung, erleichtern aber zugleich den kommunikativen Umgang, indem sie

> den Kommunizierenden mehr oder weniger feste Orientierungen für die Produktion und Rezeption von Texten geben.“ (Brinker, 1985, S. 124)

Trotz fehlender konkreter(er) Definition von *Textsorte*, handelt es sich bei Textsorten – so der kleinste gemeinsame Nenner – um Klassen von Texten, denen bestimmte Merkmale gemeinsam sind. Die Frage ist jedoch nun, welche diese textsortenbestimmenden Merkmale sind. In den verschiedenen Ansätzen werden oftmals verschiedene Merkmale vermischt, „innersprachliche, funktionale und situative“ (Vater 2001, S. 157), weshalb sich eine Vielzahl an Textsortenbegriffen wiederfinden. Isenberg (1983) bemängelt diesen Zustand und führt dies auf das fehlende empirische Fundament zurück. Seinerzeit wurde jedoch weniger ein psychologischer Ansatz angestrebt, der die psychologische Realität bestimmter Textsorteneigenschaften untersucht und diese als Basis der Textklassifikation heranzieht, als vielmehr korpusbasierte, gesprächs- und konversationsanalytische Verfahren (vgl. W. Heinemann & Viehweger, 1991).

Erschwerend für die Textsortendefinition und Klassifikation kommt hinzu, dass häufig ein Textbegriff gewählt wird, der gesprochensprachliche Kommunikations- und Darstellungsformen miteinschließt.

Nach Vater (2001, S. 159) lassen sich alle bisherigen Textklassifikationen in drei Muster einordnen:

- Klassifikation nach Gegenstand und Zielsetzung
- Klassifikation nach in Texten vorkommenden Typen von Teiltexten
- Klassifikation nach kommunikationsorientierten Kriterien

Gansel & Jürgens (2009, S. 56) schlagen fünf einfachere Unterscheidungskriterien vor:

- Textfunktion
- Verfahren zur Vertextung des Themas
- Kommunikationssituation
- Medium
- Textinhalt

Schank und Schoenthal (1976) betonen ergänzend situative Merkmale wie *Anzahl der Kommunikationsteilnehmer*, *Verhältnis der Kommunikationsteilnehmer zueinander*, *Bekanntheitsgrad*, *Öffentlichkeitsgrad* etc.

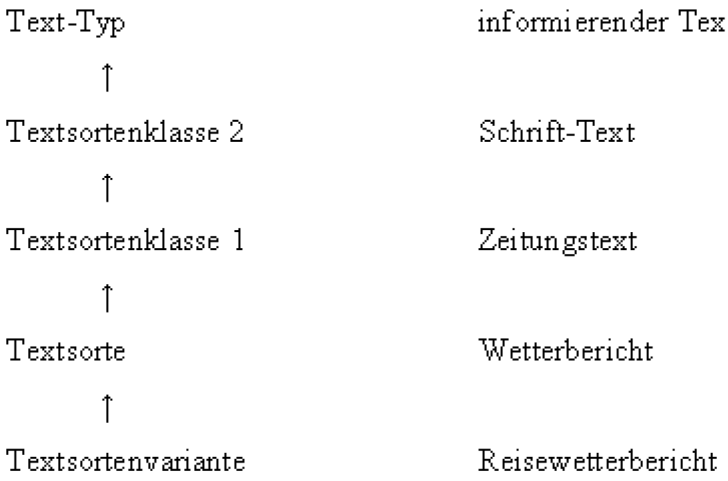
Helbig (1975, S. 73) listet folgende Textsortenkriterien:

- monologisch – dialogisch
- spontan – nicht spontan (Subkategorisierung: gedanklich (nicht) vorgeformt, sprachlich (nicht) vorher fixiert)
- Partner präsent / absent
- Zahl der Sender und Empfänger
- Öffentlichkeit
- Spezifikation der Sprechpartner (Zugehörigkeit zu bestimmten gesellschaftlichen Gruppen)
- gesprochen – geschrieben
- Modalität der Themenbehandlung (argumentativ, deskriptiv etc.)
- Grad der Steuerung bzw. des kommunikativ-theoretischen Aufwandes

Riesel und Schendels (1975) ergänzen bzw. querklassifizieren mit einer Unterscheidung von fünf Funktionalstilen (*öffentliche Rede*, *Wissenschaft*, *Presse/Publizistik*, *Alltagsrede*, *schöne Literatur*).

Nicht nur hinsichtlich Art und Umfang von textsortenbestimmenden Merkmalen besteht bislang kein Konsens, auch die Frage nach dem Auflösungsgrad, d. h. die Abstraktionsebene, auf welcher Textsorten anzusiedeln sind (z. B.: *Brief* vs. *persönlicher Brief*, *öffentlicher Brief* etc. vs. *Leserbrief*, *Liebesbrief*, *Bewerbungs(an)schreiben* etc.), ist theoretisch und empirisch ungeklärt. W. Heinemann (2000b) stellt hierzu ein hierarchisches Modell vor, welches am folgenden Beispiel (Abbildung 6.2.1) aus M. Heinemann und Heinemann (2002, S. 143) illustriert wird:

Abbildung 6.2.1: ***Hierarchisches Textklassifikationsmodell nach Heinemann (2000b).***



Was Heinemann im Rahmen dieses Schemas *Text-Typ* nennt, bezeichnet er in anderen Schriften *Textmuster, Textstrukturmuster* oder *Vertextungsmuster* (W. Heinemann, 2000a, 2008, 2009). Dies entspricht den *Verfahren zur Vertextung des Themas* von Gansel und Jürgens (2009), der *Modalität der Themenbehandlung* von Helbig (1975) und dem in Augst et al. (2007) verwendeten Begriff der *Textsorte*. Weitere Ausdrücke für diese Ebene sind unter anderem *Organisationsmuster* (Püschel, 2000), *Diskursformen* (Fienemann, 2006; Rehbein 1988), *Textformen* (Becker-Mrotzek & Böttcher, 2014), *Schreib-* oder *Kommunikationsformen* (Bucher, 1999; Budde, 2012).[30] Im Rahmen dieser Arbeit (mit Ausnahme der Kapitel 6.1. und 6.2., welche sich an der Terminologie der zitierten Autoren orientieren) wird diese Ebene der Textklassifikation, welche auch der Normierungsstudie und den entsprechenden Kompetenzstufenmodellen zugrunde liegt, im Anschluss an Heinemann und in Konsistenz u. a. mit BIFIE (2012), M. Fix (2006), Merz-Grötsch (2005; 2010; auch: *Vertextungsstrategien*) oder Richter (2008), *Textmuster* genannt. Es sei jedoch darauf hingewiesen, dass dieser Terminus von anderen Autoren anders gebraucht wird (vgl. z. B. U. Fix, 2008b).[31]

[30] Im Englischen werden prädominant die Begriffe *discourse mode*, *text genre* und *text type* verwendet.

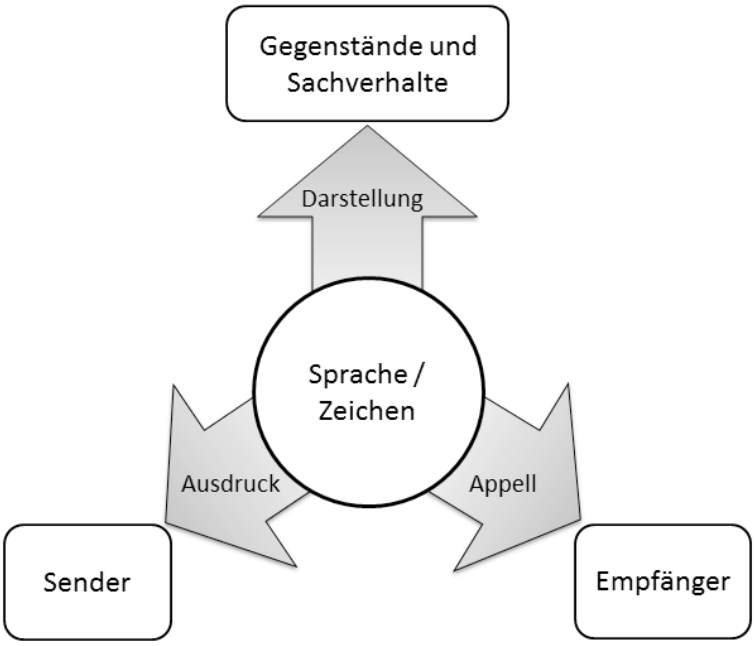
[31] Unabhängig von den Bezeichnungen, ist die Art dieser Textklassifikation nach Diskursmodi unumstritten. Uneinigkeit herrscht jedoch auch hierbei hinsichtlich des Auflösungsgrades und der Menge an zu differenzierenden Textmustern; so wird in manchen Klassifikationen zwischen *berichten* und *beschreiben* auf höchster Klassifikationsebene unterschieden, in anderen Konzepten werden diese (Sub-)Textmuster zu *informieren* zusammengefasst. Des Weiteren findet sich bisweilen ein separater Typ *appellieren*, der in anderen Konzepten dem *argumentieren* subsummiert wird (Brinker, 1985). Eine gemeinsame Unterscheidung auf höchster Ebene, welche allen Textmusterkategorisierungen zugrunde liegt (u. a. de Beaugrande & Dressler, 1981; Feilke, 2006), ist die zwischen *informieren, argumentieren* und *erzählen*, welche auch der Normierungsstudie, der Genese der Kompetenzstufenmodelle und der vorliegenden Arbeit zugrunde gelegt wurde. Diese Dreigliederung ist auch durch die drei Hauptfunktionen der Sprache begründet, *Ausdruck*, *Darstellung* und *Appell* (vgl. die Ausführungen unter 6.3.).

Der Textsortenbegriff wird im hiesigen Rahmen auf spezifischerer Ebene angesiedelt als der Begriff des Textmusters, so etwa im Rahmen der semiholistischen Stilskala, auf welcher zutreffende Textsorten, textsortenspezifische Elemente oder textsortenspezifischer Sprachgebrauch bewertet wird. Die an dieser Stelle definierten Textsorten (z. B. Lexikoneintrag, Bewerbungsschreiben, Leserbrief) entspricht dem Textsortenbegriff im obigen Schema von M. Heinemann und Heinemann (2002). Der Ausdruck *textsorten-spezifische Merkmale* bezieht sich im Rahmen dieser Arbeit auf Merkmale, die der Textsorte, Textsortenklasse oder Textsortenvariante nach Heinemann und Heinemann (2002) zuzuordnen sind und umfasst auch die oben gelisteten Merkmale nach Gansel & Jürgens (2009), Helbig (1975) sowie Schank und Schoenthal (1976).

6.3. Textmusterspezifische Anforderungen

Unterschiedliche Textmuster sind mit unterschiedlichen Anforderungen an den Schreiber verbunden (Feilke, 2006). Diese differenten Anforderungen beruhen auf den unterschiedlichen Funktionen der jeweiligen Texte der entsprechenden Muster. Generell erfüllt Sprache drei Hauptfunktionen, *Ausdruck*, *Darstellung* und *Appell*. Karl Bühler veranschaulichte dies im Rahmen des *Organon-Modells* (Bühler, 1934; vgl. Abbildung 6.3.1). Auch wenn sprachlichen Zeichen und Gebilden prinzipiell alle drei Funktionen zugrunde liegen, so fokussieren die drei Haupttextmuster jeweils eine dieser Funktionen. Beim Erzählen dominiert die sender- bzw. autororientierte Ausdrucksfunktion, beim Argumentieren die empfänger- bzw. leserausgerichtete Appellfunktion, beim Informieren die objekt- und sachverhaltsbezogene Darstellungsfunktion (Haueis, 2006).

Abbildung 6.3.1: ***Organon-Modell nach Karl Bühler (angepasste Nachbildung).***



Textmusterspezifische Anforderungen zeigen sich auf unterschiedlichen kommunikativen und sprachlichen Ebenen. So ist etwa beim Erzählen die primäre Aufgabe eine Kontextualisierung (Ohlhus 2014) vorzunehmen, welchen den Leser in das Setting des Erzählten einführt. Im Fortgang der Erzählung, welche global einer chronologischen Abfolge entspricht, folgt die Darstellung eines oder mehrerer Ereignisse, die *Komplikation*, bis hin zu einem Höhe- oder Wendepunkt, und schließt mit einer Coda (Merz-Grötsch, 2010; Gülich & Hausendorf, 2000). Dabei ist primäre Funktion des Erzählens, den Leser zu unterhalten. Hierfür werden narrative sprachliche Mittel eingesetzt, welche Spannung erzeugen und emotionale Involviertheit etablieren. Die Erzählperspektive ist eine subjektive (Gülich & Hausendorf, 2000; Quasthoff, 1993).

Anders verhält sich dies für informierende Texte. Hier wird auf etwas außerweltlich Gegebenes referiert. Die Darstellung dieses Gegebenen erfolgt sachlich und objektiv mit Anspruch auf Wahrheitstreue. Aus diesem Grund werden informierende Textformen auch *Repräsentativa* (Brinker, 1985) genannt. Die Abfolge der Ereignisse folgt hierbei unterschiedlichen Prinzipien, so erfolgt für Beschreibungen, vor allem Vorgangs-beschreibungen, eine chronologische Orientierung. Für andere Formen der Beschreibungen (bspw. lokale Beschreibungen) „wird eine Chronologie in den Raum projiziert, etwa als Bewegung durch den Raum. Rehbein (1984, [S.] 79) spricht hier von einem Gang durch den Vorstellungsraum“ (Becker-Mrotzek & Böttcher, 2014, S. 148). Hierbei steht der Schreiber

vor der Herausforderung, dem Leser durch räumliche und zeitliche deiktische sprachliche Mittel eine präzise Orientierung zu bieten (Ossner, 2014). Auch für das Berichten obliegen dem Schreiber die Gebote der Sachlichkeit, Objektivität und Präzision (Ludger Hoffmann, 1996). Berichte erfolgen resultatorientiert (Feilke, 2014), folgen in ihrer Strukturierung gemäß der Priorität der Informationen und orientieren sich an der Beantwortung der sogenannten W-Fragen (Was geschah? – Wer war beteiligt? – Wo und wann ereignete sich das Geschehen? – Wie geschah es? – Was waren die Ursachen? – Welche Folgen gibt/gab es?) (M. Fix, 2006; Merz-Grötsch, 2010).

Beim Argumentieren ist die primäre Funktion des Textes, den Leser zu überzeugen bzw. eine Problemlage dem Leser nahezubringen; je nach Ziel liegt eine stärkere Partner- und Situations- oder Inhalts- und Problemorientierung vor (Friedrich, 1988; Pohl, 2014). Die vier Hauptanforderungen eines argumentierenden Textes sind nach Schneider und Tetling (2014) das adäquate Erfassen der Sache, die hinreichend richtig antizipierte Perspektive des Lesers, die Textgestaltung sowie die Verdeutlichung des eigenen Standpunkts unter einer dennoch objektiven Darstellung. Die Textstruktur folgt hierbei der groben Struktur der Themeneinführung, der Argumentation und der finalen Schlussfolgerung, des Fazits. Im Kernteil der Argumentation stehen Argumente, die selbst eine interne Struktur von Prämissen und Konklusion unterliegen (Bayer, 2007; Ossner, 2006; Toulmin, 1958; Völzing, 1980). Die Abfolge der Argumente kann sich hierbei an verschiedenen Gewichtungsprinzipien wie bspw. dem Pyramidenprinzip (vom schwächsten zum stärksten Argument) oder dem Sanduhrprinzip (zu Beginn starke Argumente, zur Mitte schwächere Argumente, gegen Ende wiederum starke Argumente) orientieren. Bei dialogischen Argumentationen sind darüber hinaus Abwägungen zwischen den einzelnen Positionen vorzunehmen und diese inhaltlich und sprachlich sinnvoll zueinander in Beziehung zu setzen (Pohl, 2007, 2014).

Die hier angeführten Anforderungen an das Schreiben von Texten bestimmter Textmuster kann lediglich als exemplarisch und auf wichtige Aspekte fokussiert angesehen werden. In Anlehnung an Becker-Mrotzek und Böttcher (2014, S. 90), M. Fix (2006, S. 106–107) sowie Merz-Grötsch (2010, S. 123, 146, 164, 206) findet sich in Tabelle 6.3.1 eine Gegenüberstellung verschiedener Aspekte der Textmuster *Erzählen*, *Argumentieren* sowie der informierenden Textmuster *Berichten* und *Beschreiben*.

Tabelle 6.3.1: ***Prototypische textmusterspezifische Anforderungen im Vergleich: Erzählen, Berichten, Beschreiben, Argumentieren.***

	Erzählen	**Berichten**	**Beschreiben**	**Argumentieren**
Schreibziel/ Funktion	Ein Ereignis (erlebt, rezipiert oder fiktiv) wird unterhaltend, spannend und leserinvolvierend dem Adressaten nahegebracht werden.	Es wird vom Resultat her über ein Ereignis informiert.	Ein Sachverhalt, Objekt oder Prozess wird zum Nachvollzug / zur Wiedererkennung einem Adressaten dargestellt.	Ein Problem wird beleuchtet und sich mit diesem auseinandergesetzt, um eine eigene Stellungnahme zu begründen und den Leser zu überzeugen.
Modalität der Themen-entfaltung	subjekt-orientiert, emotional, szenisch, perspektivisch, effekt-orientiert	sachlich, informierend, referierend, deskriptiv, präzise, zweckbezogen	sachlich, informierend, übersichtlich, deskriptiv, präzise, objekt- oder prozessbezogen, begrifflich eindeutig, anschaulich	adressaten-orientiert, persuasiv, problembezogen
(Typische) sprachliche Mittel	Tempus: Präteritum, wörtliche Rede, Ich-Form, abwechslungsreiche Satzmuster	Tempus: Präteritum, unpersönliche Sprachformen (Passiv), indirekte Rede, Fachsprache	Tempus: Präsens, unpersönliche Sprachformen (Passiv), deiktische Adverbien, exakte Wortwahl, Fachbegriffe, reihendes Satzgefüge	Tempus: Präsens, dialogische Strukturen, Konnektoren, Zitate und Referenzen, objektive Sprachhaltung, präzise Sprachverwendung
Textstruktur	Orientierung (Exposition) – Komplikation – Höhepunkt – Auflösung – Coda	Logische, resultat-orientierte Ordnung der Ereignisse (W-Fragen), zweckbezogene Selektion	Orientierung an der Sach- bzw. Prozessstruktur (Gang durch den Vorstellungsraum)	Einleitung mit Fragestellung – Darstellung der Standpunkte – Anführen von Argumenten – Schlussfolgerung(en), Begründung der eigenen Position
Schreibplan, Wissens-organisation	Ideensammlung, Brainstorming, Imagination, Mindmap, Erinnern	Ziel klären, W-Fragen beantworten, zweckbezogen selektieren	genaue Beobachtung, Prozesse zeitlich gliedern, Gegenstände räumlich ordnen, begrifflich-kategoriale Benennungen	Thema und Problem klären, Stoffsammlung/ Recherche, Argumente anordnen, Belege und Beispiele suchen

6.4. Fragestellungen und Hypothesen

Leitfragestellung der hier vorgestellten Untersuchung ist, ob Schreibkompetenzen textmusterspezifisch bzw. textmusterunabhängig sind.

Gemäß den unterschiedlichen textmusterspezifischen Anforderungen und im Anschluss an bisherige Befunde – auch wenn diese nur in wenigen Fällen auf systematischen Erhebungen basieren – ist Textmusterspezifität von Schreibkompetenzen zu erwarten.

Darüber hinaus soll untersucht werden, ob auch für die drei Teilkompetenzen *Inhalt*, *Stil* und *sprachliche Richtigkeit* Textsortenspezifität bzw. Textsortenunabhängigkeit besteht.

Die textmusterdifferenten Anforderungen sind im inhaltlichen und stilistischen Bereich zu verorten. So wird etwa in informierenden Texten eine präzise und konkrete Wortwahl erwartet, während in erzählenden Texten der Einsatz von sprachlichen Bildern und emotionalen Markierungen den narrativen Charakter des Textes unterstreicht. Für argumentative Texte folgen Auswahl und Arrangement der Inhalte idealiter nach Gewichtungsprinzipien, während erzählenden Texten in der Regel eine chronologische und/oder kausale Abfolge zugrunde liegt. Aus diesem Grund der Textmusterspezifität inhaltlicher und stilistischer Anforderungen bei dem Schreiben von Texten, wird auch für die Teilkompetenzen *Inhalt* und *Stil* Textmusterspezifität antizipiert. Da orthografische und grammatische Regeln unabhängig von Textmustern und Textsorten Gültigkeit besitzen, sollten diese Regeln über Textmuster hinweg beherrscht (bzw. nicht beherrscht) werden; für die Domäne *sprachliche Richtigkeit* wird somit Textmusterunabhängigkeit erwartet.

Im Rahmen der Analysen wird auch untersucht, ob die innere Struktur von Schreibkompetenz, d. h. das Verhältnis der Dimensionen *Inhalt*, *Stil* und *sprachliche Richtigkeit* zueinander textmusterübergreifend konstant ist oder über Textmuster variiert. Da sich hierzu theoretisch keine Hypothesen aus den Anforderungen an das Schreiben von Texten bestimmter Textmuster ableiten lassen und bisher keine textmusterkontrastierende Untersuchungen zur internen Struktur von Schreibkompetenzen vorliegen, wird dieser Aspekt explorativ untersucht. Es sei angemerkt, dass aus der Annahme eines textsortenspezifischen Konstrukts *Schreibkompetenz* sowie textmusterspezifischer inhaltlicher und stilistischer Schreibteilkompetenzen nicht folgt, dass sich auch die interne Struktur von *Schreibkompetenz* textmusterspezifisch unterscheidet; unterscheidet sich allerdings die interne Struktur von *Schreibkompetenz* über Textmuster hinweg, wäre diese Differenz in der inneren Struktur ein weiterer Beleg für die Textmusterspezifität des Konstrukts.

6.5. Durchführung der Studie

6.5.1. Datengrundlage

Als Datengrundlage dienten die anhand der holistischen und semiholistischen Skalen ermittelten Schreibleistungsdaten der 2996 Schülerinnen und Schüler der Normierungsstudie (vgl. Kapitel 3.4.). Diese Schülerinnen und Schüler bearbeiteten je zwei Schreibaufgaben. Je circa ein Viertel der Schülerschaft bearbeitete zwei argumentierende (n = 689), zwei informierende (n = 594) bzw. zwei narrative Aufgaben (n = 668). Für das übrige Viertel (n = 632) lagen zwei Aufgaben unterschiedlicher Textmuster vor, dabei waren alle drei Textmusterkombinationen zu etwa gleichen Teilen vertreten (n = 202; 206; 224).

6.5.2. Analysen

Um zu prüfen, ob die Schreibleistungen der Schülerinnen und Schüler textmusterunabhängig oder textmusterspezifisch sind, wurden auf manifester Ebene die jeweils beiden Aufgabenbearbeitungen eines Schülers bzw. einer Schülerin herangezogen und die Spearman-Korrelationen bzw. die entsprechenden Fisher-z-Werte zwischen den aufgabenspezifischen Leistungen verglichen (Fisher, 1921; Myers & Sirois, 2006; Raghunathan, Rosenthal & Rubin, 1996; Steiger, 1980). Ziel dieser Analyse war es, zu prüfen, ob diese Korrelationen für Schülerinnen und Schüler, welche zwei Aufgaben desselben Textmusters bearbeiteten, bedeutsam höher sind als für Schülerinnen und Schüler, welche zwei Aufgaben bearbeiteten, die unterschiedlichen Textmustern zuzuordnen sind.

Zur Überprüfung analoger Fragestellung auf latenter Ebene wurde ein Modellvergleich durchgeführt: Das Konstrukt *Schreibkompetenz* wurde hierfür im Rahmen zweier IRT-Modelle (ordinale Rasch-Modelle, vgl. Kapitel 3.6.), i. e. zum einen eindimensional, zum anderen dreidimensional (gemäß der drei Textmuster) auf Basis der Globalurteile, mit welchen die Texte bewertet wurden, modelliert. Die Modellierungen wurden mit *ConQuest* (Version 2.0) (Wu et al. 1998) durchgeführt. Anschließend wurden die Modelle hinsichtlich ihrer Passung/Güte miteinander verglichen. Dieser Vergleich erfolgte über einen χ^2-Test der Differenz der *Deviance*-Werte unter Einbeziehung der geschätzten Parameter der beiden Modelle (Adams & Wu, 2010; Winther, 2010).

Zur Untersuchung der inneren Struktur von Schreibkompetenz wurden die Schreibleistungsdaten anhand der semiholistischen Subskalen (*Inhalt*, *Stil* und *sprachliche Richtigkeit*) herangezogen und einerseits textmusterunabhängig in einem dreidimensionalen (je eine Dimension für inhaltliche, stilistische und orthografisch-grammatische Schreibkompetenz), zum anderen textmusterdifferenzierend in einem neundimensionalen ordinalen Rasch-Modell (je eine Dimension für inhaltliche, stilistische und orthografisch-grammatische Schreibkompetenz × je eine Dimension für die Textmuster *argumentieren*, *informieren* und *narrativ*) skaliert. Drei Dinge waren hierbei von Interesse:

(i) Erweist sich die textmusterspezifische Modellierung der inneren Struktur hinsichtlich der Modellpassung als angemessener als die textmusterunabhängige? Dies wurde wiederum anhand eines χ^2-Tests der Differenz der *Deviance*-Werte der beiden kontrastierten Modelle geprüft.

(ii) Falls sich die innere Struktur textmusterspezifisch unterscheidet, inwiefern, d. h. hinsichtlich welcher Relationen unterscheiden sich die entsprechenden Strukturen? Hierzu wurden die Konfidenzintervalle für die Korrelationen zwischen den Konstrukten bestimmt, anhand derer die Korrelationen verglichen werden konnten (Brandstätter, 1999; Gardner & Altman, 1986; Wang 2004). Zur Ermittlung der Konfidenzintervalle wurden entsprechende IRT-Modellierungen mit der Software *Mplus* (Version 5.21) wiederholt (Muthén & Muthén, 1998, 2008).[32]

(iii) Erweisen sich die einzelnen textmusterspezifisch ermittelten Teilkompetenzen (*Inhalt*, *Stil* und *sprachliche Richtigkeit*) als hinreichend hoch miteinander korreliert, so dass sie sich als dasselbe psychologische Konstrukt interpretieren lassen? Handelt es sich beispielsweise bei der inhaltlichen Schreibkompetenz im Rahmen argumentierenden Schreibens um dieselbe Teilkompetenz wie bei der inhaltlichen Schreibkompetenz im Rahmen informierenden Schreibens oder stellen sie verschiedene Teilkompetenzen dar?

[32] Die Wiederholung der Analyse war dem Faktum geschuldet, dass *ConQuest* keine Standardfehler und/oder Konfidenzintervalle für die latenten Korrelationen liefert. Prinzipiell wäre eine Berechnung aller Analysen ausschließlich in *Mplus* realisierbar gewesen, aufgrund der Anbindung an und Vergleichbarkeit mit den Analysen zur Auswertung der Normierungsstudie im Rahmen der Genese der Kompetenzstufenmodelle (vgl. Kapitel 3 und 4) stützen sich die Hauptanalysen auch in diesem Rahmen auf eine Modellierung mit *ConQuest*. Für die Analysen in *Mplus* wurden die Faktorenladungen eines Konstrukts gleichgesetzt, sodass ein Rasch-Modell imitiert wurde – Skalierungen in *Mplus* beruhen bei Nichtrestriktion auf einem 2-paramametrischen logistischen Verfahren, während das Rasch-Modell, wie durch *ConQuest* modelliert, ein 1-parametrisches logistisches Modell ist.

6.6. Ergebnisse

Die Berechnung der manifesten Korrelationen zwischen den beiden aufgabenspezifischen Schreibleistungswerten nach Aufgaben- und Textmusterkombinationen führte zu den in den Tabellen 6.6.1, 6.6.2. und 6.6.3. angeführten Ergebnissen.

Tabelle 6.6.1: ***Spearman-Korrelationen zwischen aufgabenspezifischen manifesten Schreibleistungswerten von Schülerinnen und Schülern bei der Bearbeitung zweier Schreibaufgaben nach Textmustern.***

	A-A	I-I	N-N	A-I	A-N	I-N
Aufgabenebene (Werte je Aufgabenkombination)	.50 (243)	.43 (193)	.25 (229)	.12 (202)	.39 (224)	.36 (206)
	.43 (219)	.43 (185)	.24 (223)			
	.33 (227)	.24 (216)	.14 (216)			

A: argumentierende Aufgabe; I: informierende Aufgabe; N: narrative Aufgabe; Werte in Klammern: Anzahl der Fälle (*n*); es lagen jeweils drei Aufgabenkombinationen innerhalb eines jeden Textmusters vor, nur jeweils eine Aufgabenkombination, welche je zwei Textmuster kombinierte.

Tabelle 6.6.2: ***Spearman-Korrelationen zwischen manifesten Schreibleistungswerten von Schülerinnen und Schülern bei der Bearbeitung zweier Schreibaufgaben nach Textmustern (über Aufgabenkombinationen hinweg).***

	A-A	I-I	N-N	A-I	A-N	I-N
Textmusterebene (aufgabenübergreifend)	.45(689)	.31 (594)	.21 (668)	.12 (202)	.39 (224)	.36 (206)

A: argumentierende Aufgabe; I: informierende Aufgabe; N: narrative Aufgabe; Werte in Klammern: Anzahl der Fälle (*n*); für die textmusterdifferenten Kombinationen (A-I, A-N, I-N) sind aufgrund des Vorliegens nur einer Aufgabenkombination für diese Textmusterkombinationen die Werte identisch mit den aufgabenspezifischen Werten in Tabelle 6.6.1.

Tabelle 6.6.3: ***Spearman-Korrelationen zwischen manifesten Schreibleistungswerten von Schülerinnen und Schülern bei der Bearbeitung zweier Schreibaufgaben nach Textmusteridentität vs. Textmusterdiversität.***

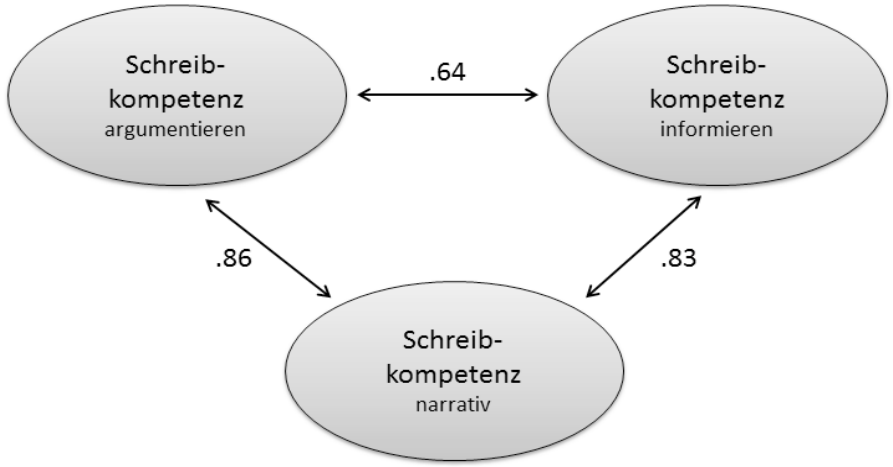
Aufgaben identischer Textmuster	Aufgaben unterschiedlicher Textmuster
.32 (1951)	.29 (632)

Werte in Klammern: Anzahl der Fälle (*n*)

Betrachtet man zunächst lediglich die argumentierenden und informierenden Aufgaben bzw. Aufgabenkombinationen, zeigen sich deutlich höhere Leistungszusammenhänge für die Bearbeitung textmusteridenter Aufgaben als für textmusterdifferente Aufgaben (für beide Fisher-*z*-Vergleiche $p < .01$). Nicht nur aufgabenübergreifend (Tabelle 6.6.2.), sondern auch für jedes einzelne textmusteridente Aufgabenpaar (Tabelle 6.6.1.) erweisen sich die Korrelationen als höher gegenüber der Korrelation der Leistungen bei der Bearbeitung einer argumentierenden und einer informierenden Aufgabe (für alle der Fisher-*z*-Vergleiche $p < .05$). Dieses Muster ändert sich jedoch, sobald man auch die Narration betrachtet. Hier zeigt sich zum einen kein besonders (mit den anderen Textmuster vergleichbar) starker Zusammenhang bei der Bearbeitung zweier textmusteridenter Aufgaben. Zum anderen erweisen sich die Zusammenhänge bei der Bearbeitung einer narrativen und einer argumentierenden Aufgabe sowie bei der Bearbeitung einer narrativen und einer informierenden Aufgabe als höher als bei der Bearbeitung zweier narrativer Aufgaben (alle $p < .05$). In der Gesamtschau zeigt sich unter Einbeziehung aller zwölf Aufgaben, aller zwölf real vorliegenden Aufgabenkombinationen und aller drei Textmuster ein ähnlich hoher Leistungszusammenhang ($p = .24$) für die Bearbeitung textmusteridenter Aufgaben ($r = .32$) wie für die Bearbeitung textmusterdifferenter Aufgaben ($r = .29$).

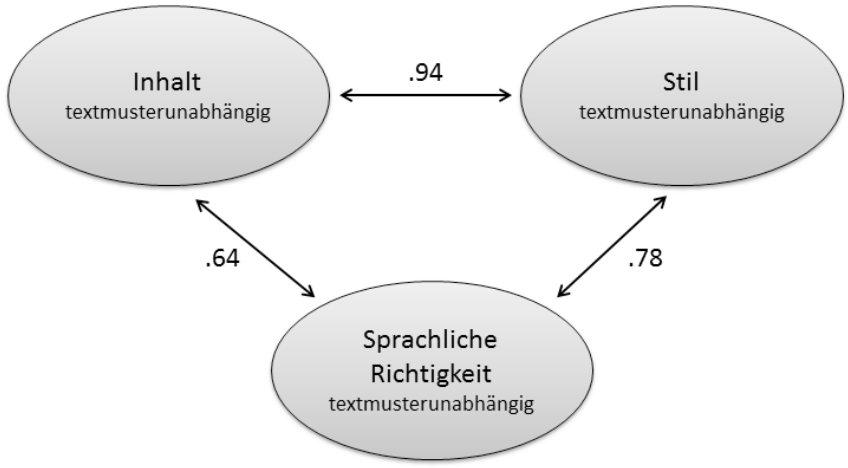
Die IRT-Modellierungen von *Schreibkompetenz* anhand der Globalurteile weisen für das eindimensionale Modell eine *Deviance* von 14585.71 bei 49 frei geschätzten Parametern auf; für das dreidimensionale Modell beträgt die *Deviance* 14571.62 bei 54 geschätzten Parametern. Der χ^2-Test auf Basis der *Deviance*-Differenz (14.09) bei 5 (= 54 − 49) Freiheitsgraden erweist sich als statistisch bedeutsam ($p = .015$). Das dreidimensionale Modell ist somit angemessener. Die bei dreidimensionaler Modellierung ermittelten Zusammenhänge auf latenter Ebene zwischen den textmusterspezifischen Schreibkompetenzen sind in Abbildung 6.6.1 dargestellt.

Abbildung 6.6.1: ***Textmusterspezifische (dreidimensionale) Modellierung von Schreibkompetenz: Latente Zusammenhänge zwischen den textmusterspezifischen Konstrukten.***



Zwischen dem informierenden und argumentierenden Textmuster zeigt sich eine moderate Korrelation zwischen den latenten Konstrukten (r = .64). Beide Konstrukte weisen jedoch einen höheren Zusammenhang (>.80) mit dem narrativen Schreibkompetenzkonstrukt auf.

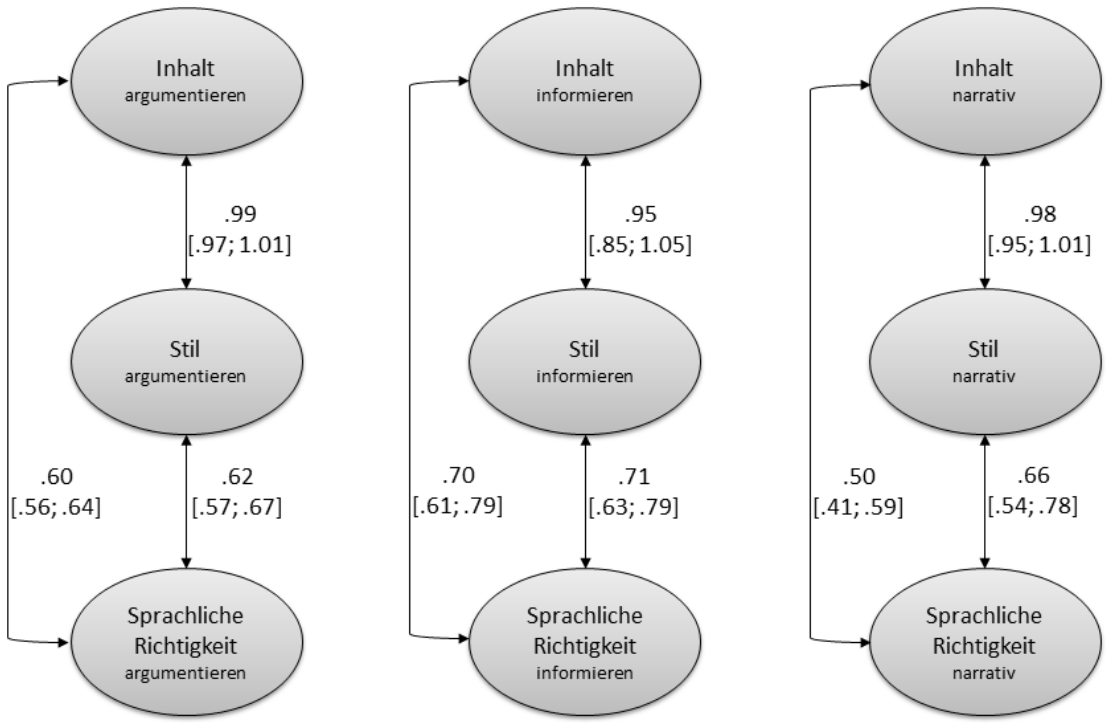
Abbildung 6.6.2: ***Latente Zusammenhänge zwischen den Fähigkeitsdimensionen Inhalt, Stil und sprachliche Richtigkeit bei textmusterunabhängiger (dreidimensionaler) Modellierung.***



Hinsichtlich der Modellierung der internen Struktur von *Schreibkompetenz* erweist sich die textmusterdifferenzierende neundimensionale Modellierung (vgl. Abbildung 6.6.3.) als geeigneter als die textmusterunabhängige dreidimensionale Skalierung (vgl. Abbildung 6.6.2). Die *Deviance* für das neundimensionale Modell beträgt 23618.15 bei 153 geschätzten

Parametern, die *Deviance* für das dreidimensionale 23710.01 bei 114 geschätzten Parametern; der entsprechende χ^2-Tests auf Basis der *Deviance*-Differenz (91.87) bei 39 Freiheitsgraden erweist sich als statistisch signifikant ($p < .001$). Dabei zeigt sich jedoch ein weitgehend homogenes Bild hinsichtlich der Dimensionalität von *Schreibkompetenz*. Es besteht für alle drei Textmuster ein sehr enger Zusammenhang ($r > .90$) zwischen inhaltlicher und stilistischer Schreibkompetenz (vgl. Abbildung 6.6.3), die 95%-Konfidenzintervalle der entsprechenden Korrelationen schließen die 1 jeweils ein; dies spricht dafür, dass es sich bei den beiden Faktoren *Inhalt* und *Stil* um dasselbe Konstrukt, zumindest jedoch um zwei sehr eng verbundene Konstrukte handelt. Geprüft wurde dies zusätzlich, indem ein sechsdimensionales Modell (*Inhalt+Stil* als eine, *sprachliche Richtigkeit* als zweite Dimension × drei Textmusterdimensionen) mit dem neundimensionalen verglichen wurde, es zeigten sich keine bedeutsamen Unterschiede zwischen den beiden Modellen (Devianzdifferenzvergleich: $\Delta = 21.4$; $df = 24$; $p = .614$).

Abbildung 6.6.3: ***Latente Zusammenhänge zwischen den Fähigkeitsdimensionen Inhalt, Stil und sprachliche Richtigkeit bei textmusterspezifischer (neundimensionaler) Modellierung.***

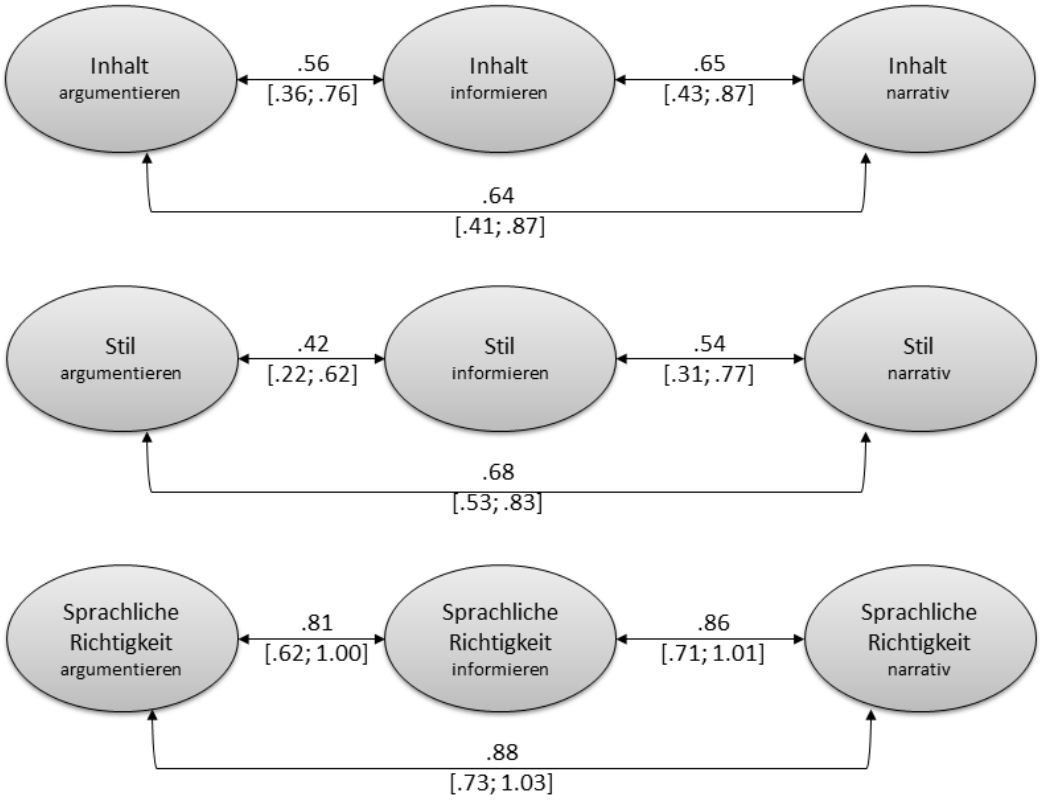


Klammerangaben: 95%-Konfidenzintervalle[33]

[33] Korrelationen > 1 beruhen auf der softwarebasierten Schätzung der Konfidenzintervalle. Da konkrete Korrelationen auf einen Wertebereich zwischen -1 und 1 beschränkt sind, sind diese Korrelationswerte > 1 als 1 zu interpretieren.

Deutlich niedriger als die Korrelationen zwischen *Inhalt* und *Stil* fallen die Korrelationen zwischen je einer dieser beiden Dimensionen und der Dimension der sprachlichen Richtigkeit aus (vgl. Abbildung 6.6.3). Dabei unterscheiden sich die Textmuster voneinander mit einem bedeutsam höheren Zusammenhang zwischen *Stil* und *sprachliche Richtigkeit* für *informieren* als für *argumentieren,* (*narrativ* von beiden ununterschieden dazwischen,) sowie einem bedeutsam höheren Zusammenhang zwischen *Inhalt* und *sprachliche Richtigkeit* für *informieren* als für *argumentieren* als für *narrativ*.

Abbildung 6.6.4: ***Textmusterspezifische (neundimensionale) Modellierung der Teilfähigkeiten Inhalt, Stil und sprachliche Richtigkeit: Latente Zusammenhänge zwischen den textmusterspezifischen Teilfähigkeitskonstrukten.***



Klammerangaben: 95%-Konfidenzintervalle[34]

Vergleicht man die Dimensionen über die Textmuster hinweg, zeigt sich, dass die orthografisch-grammatischen Teilkompetenzen einen hohen Zusammenhang über Textmuster hinweg aufweisen ($r \geq .80$), während inhaltliche und stilistische Schreibkompetenzen lediglich moderate Korrelationen aufweisen (vgl. Abbildung 6.6.4). Die 95%-

[34] Korrelationen > 1 sind als 1 zu interpretieren (vgl. letzte Fußnote).

Konfidenzintervalle indizieren hierbei für die Dimension der sprachlichen Richtigkeit eine Nichtverschiedenheit von 1, was dafür spricht, dass es sich bei dieser um eine textmusterindifferente Dimension handelt.

6.7. Diskussion

Die Ergebnisse der Analyse der manifesten Schreibleistungsdaten liefern zunächst kein eindeutiges Bild für oder gegen die Annahme von textmusterspezifischen (vs. textmusterunabhängigen) Schreibkompetenzen. Es zeigt sich ein bedeutsamer Unterschied zwischen den Textmustern des Argumentierens und des Informierens. Für Schülerinnen und Schüler, die zwei argumentierende oder zwei informierende Aufgaben bearbeiteten, liegen höhere Leistungszusammenhänge vor als für Schülerinnen und Schüler, die eine informierende und eine argumentierende Aufgabe bearbeiteten. Auf latenter Ebene weisen die Konstrukte *argumentierende Schreibkompetenz* und *informierende Schreibkompetenz*, wie die dreidimensionale textmusterspezifische Modellierung anhand der Globalurteile zeigt, lediglich einen moderaten Zusammenhang (r = .64) auf und lassen sich somit nicht als ein- und dasselbe Konstrukt interpretieren.

Unter Einbeziehung der narrativen Schreibleistungsdaten zeigt sich ein anderes Bild. Hier zeigten sich keine höheren Zusammenhänge für Schülerinnen und Schüler, die zwei narrative Aufgaben bearbeitet hatten gegenüber Schülerinnen und Schülern, die eine narrative und eine argumentierenden oder informierende Aufgabe bearbeitet hatten. Darüber hinaus zeigten sich bei der dreidimensionalen textmusterspezifischen Modellierung anhand der Globalurteile hohe Zusammenhänge zwischen den latenten Konstrukten *narrative Schreibkompetenz* und *argumentierende Schreibkompetenz* sowie zwischen *narrative Schreibkompetenz* und *informierende Schreibkompetenz*. Vermutlich ist dieser Effekt auf die eingesetzten Aufgaben zurückzuführen. Aufgrund der niedrigen curricularen Relevanz narrativer Aufgaben am Ende der Sekundarstufe I, wurden für die Normierungsstudie für das Textmuster der Narration gegenüber den anderen Textmustern mitunter auch weniger prototypische Aufgaben entwickelt, darunter auch solche, die sich als Mischformen verstehen lassen oder bei welchen die Narration zwar im eigentlichen Fokus steht, die Rahmung aber in eine Textsorte gebettet ist, die eher einem anderen Textmuster entspricht. So musste etwa für eine der Aufgaben ein Lexikoneintrag – prototypisch eher dem informierenden Textmuster zuzuordnen – mit vorwiegend erzählenden Elementen geschrieben werden. Eine andere Aufgabe verlangte die

Erzählung einer Einzelfallgeschichte, anhand derer in einem textabschließenden Fazit eine vorgegebene These gestützt bzw. widerlegt werden sollte, womit die Rahmung der Geschichte einem argumentierenden Duktus unterliegt. Eine dritte Aufgabe erforderte die Genese eines Tagebucheintrages und ließ somit auch wertende und kommentierende Elemente zu. Nur in einer der vier Aufgaben lag die Anforderung eines ausschließlich und prototypisch narrativen Textmusters vor.[35] Gerade die beiden narrativen Aufgaben mit informierenden bzw. argumentierenden Elementen waren diejenigen Aufgaben, die im Rahmen der Spiral-Verknüpfung (vgl. Kapitel 3.4.) die Narration mit dem jeweiligen anderen entsprechenden Textmuster verbanden. Somit bestand einerseits eine höhere Diversität in den narrativen Aufgaben, andererseits eine größere Similarität zwischen Nachbaraufgaben, welche die Narration mit dem Argumentieren bzw. dem Informieren verknüpften.

Der Modellvergleich zwischen einer eindimensionalen und dreidimensionalen Modellierung von *Schreibkompetenz* auf der Basis der Globalurteile wies eine dreidimensionale und somit textmusterspezifische Modellierung als die angemessenere aus. Auch die Modellierung der aspektualen Dimensionen von *Schreibkompetenz* (*Inhalt*, *Stil*, *sprachliche Richtigkeit*) lieferte Evidenz für eine textmusterspezifische Interpretation. Neben einer bedeutsam besseren Modellpassung für das neundimensionale textmusterspezifische Modell zeigten sich für zwei der drei Schreibkompetenzdimensionen, *Inhalt* und *Stil*, lediglich moderate Zusammenhänge zwischen verschiedenen Textmustern (r = .49–.68), welche dagegen sprechen, inhaltliche und stilistische Schreibkompetenzen als textmusterunabhängige Konstrukte zu interpretieren. Lediglich für die Dimension der *sprachlichen Richtigkeit* zeigten sich höhere Zusammenhänge, die darauf hindeuten, dass diese Fähigkeiten weitgehend textmuster-unabhängig sein könnten. Auch wenn diese Korrelationen unter dem Grenzwert von .90 bleiben, ab welchem als Richtwert oftmals davon ausgeht, dass es sich um ein und dasselbe Konstrukt handele (Kline, 2011; Maehler & Schmidt-Denter, 2013; Westen & Rosenthal, 2003), so indizieren die 95%-Konfidenzintervalle, dass die Korrelationen nicht bedeutsam von 1 verschieden sind. Auch Astrid Neumann (2007) lieferte bei hohen aufgaben- und textsortenübergreifenden Parallelen in den Bereichen *Orthografie* und *Grammatik* im Rahmen detaillierterer Betrachtungen auch Evidenz für unterschiedliche orthografische und grammatische Leistungen im Vergleich verschiedener Textsorten bzw. Aufgaben mit unterschiedlichen Textsortenmerkmalen, i. e. im Kontrast formalisierter und persönlicher Briefe. Die Autorin führt dies auf die unterschiedliche Expertise im Umgang mit diesen

[35] Im Rahmen der anderen beiden Textmuster kamen ausschließlich textmuster-prototypische Aufgaben zum Einsatz.

Textsorten sowie auf unterschiedliche Anforderungen (z. B. freiere Wort- und Konstruktionswahl in persönlichen Briefen) zurück. Darüber hinaus ist anzunehmen, dass bei Schülerinnen und Schülern, die entsprechende orthografische Regeln noch nicht weitgehend internalisiert haben und die Regelbefolgung noch nicht automatisiert abläuft, aufgaben- oder textsortenspezifische Anforderungen Aufmerksamkeit sowie andere kognitive Ressourcen für deren Bewältigung binden und somit von der orthografisch-grammatischen Kontrolle abziehen könnten (Berman & Nir-Sagiv, 2007; Kellogg, 1987; McCutchen,1996). Diese Effekte der textsorten- und aufgabenspezifisch unterschiedlich starken Bindung der kognitiven Ressourcen auf Kosten der Aufmerksamkeit hinsichtlich Orthografie und Grammatik könnten somit eine plausible Ursache für die Restvarianz hinsichtlich orthografisch-grammatischer Fähigkeiten sein. Im Rahmen der hier eingesetzten Aufgaben ist es aufgrund der begrenzten Aufgabenmenge sowie des umgesetzten Designs in Form einer Spiral-Verknüpfung leider nicht möglich, auf Ebene von Textsorten oder Textsorten-merkmalen zu prüfen, wie viel gemeinsame Varianz sich auf diese Ebenen verteilt. So kamen beispielsweise in allen drei Textmustern Aufgaben zum Einsatz, die der Textsorte *Brief* zuzuordnen sind, jedoch wurde das Testdesign nicht zur Kontrolle dieser (Textsorten-)Ebene ausgerichtet, sodass die entsprechenden Aufgaben nicht unmittelbar miteinander verknüpft waren, d. h. keine Schülerin und kein Schüler bearbeitete zwei textsortenidentische, aber textmusterdifferente Aufgaben. Auch lag keine aufgabenspezifische thematische Kontrolle der eingesetzten Aufgaben wie zumindest basal in der Studie von Olinghouse et al. (2012) realisiert vor, sodass auch mögliche Varianzanteile auf thematischer Aufgabenebene nicht systematisch aufgeklärt werden können.

Bezüglich der inneren dimensionalen Struktur von *Schreibkompetenz* zeigte sich textmuster-übergreifend ein Befundmuster, welches eine zweidimensionale Struktur nahelegt. Inhaltliche und stilistische Schreibkompetenzen erweisen sich als sehr hoch miteinander assoziiert ($r >$ .90), ein Vergleich der Modellierung der beiden Faktoren als ein Konstrukt zeigte keine statistisch schlechtere Modellpassung als die zweifaktorielle Modellierung, sodass hierbei von einem Konstrukt ausgegangen werden kann. Die Dimension der *sprachlichen Richtigkeit* erweist sich deutlich niedriger mit *Inhalt* und *Stil* assoziiert, weshalb von einer eigenständigen Teilfähigkeit auszugehen ist. Dies steht im Einklang mit bisherigen Befunden zur inneren Struktur von *Schreibkompetenz* (vgl. Kapitel 2.3.5.): Astrid Neumann (2007) fand in ihren Untersuchungen Evidenzen für eine zweidimensionale Struktur mit den Dimensionen *Semantik/Pragmatik*, unter welche ausschließlich Merkmale erfasst werden, die hier den Dimensionen *Inhalt* oder *Stil* zuzuordnen sind, sowie *Sprachsystem*, worunter weitestgehend

Merkmale der sprachlichen Richtigkeit gefasst werden. Böhme et al. (2009), die ein zu dieser Studie analoges Auswertungssystem an holistischen und semiholistischen Skalen verwendeten, konnten für die Primarstufe ebenfalls Evidenz für eine zweidimensionale Struktur mit *Inhalt+Stil* als eine und *sprachliche Richtigkeit* als andere Dimension erbringen, auch wenn die Korrelationen zwischen den Dimensionen *Stil* und *Inhalt* etwas geringer ausfielen (r = .88; S. 319), was jedoch im Einklang steht mit einer angenommenen konvergenten Entwicklung sprachlicher Fähigkeiten bis zum Ende der Sekundarstufe I (Diakidoy, Stylianou, Karefillidou & Papageorgiou, 2005; Fitzgerald & Shanahan, 2000; Ginther & Stevens, 1998; Heller, 1999; Jude, 2008; J. Langer & Flihan, 2000).[36]

Hinsichtlich der textmusterspezifischen Unterschiede in der Höhe der Zusammenhänge zwischen den einzelnen Dimensionen werden wiederum aufgaben- und textsortenspezifische Effekte vermutet. So lassen sich etwa informierende Aufgaben nur sinnvoll auswerten, wenn sie relativ strikte Vorgaben machen (vgl. auch Kapitel 7.1. zur generellen Notwendigkeit an Vorgaben für Testaufgaben). Informierende Aufgaben fordern per se ein hohes Maß an Präzision und stellen somit hohe Anforderungen an die Lexik sowie an die Strukturierung und das Arrangement der Inhalte (vgl. A. Neumann, 2007); im Rahmen der Engführung der Aufgaben zu Testzwecken werden die ohnehin schon eingeschränkten Realisierungs-möglichkeiten nochmals restringiert. So kamen auch in der Normierungsstudie für das informierende Textmuster stilistisch anspruchsvolle und in den Antwortmöglichkeiten wenig variable Aufgaben zum Einsatz. Auch hier stellt die Bildung der kognitiven Ressourcen einen plausiblen Erklärungsfaktor für die erhöhte Korrelation zwischen *Stil* bzw. *Inhalt* einerseits und *sprachliche Richtigkeit* andererseits im informierenden Textmuster dar (McCutchen, 1996). Zur Bewältigung der strengen stilistischen und inhaltlichen Anforderungen werden kognitive Ressourcen (bspw. Aufmerksamkeit) gebunden, die für andere Kontrollebenen nicht mehr zur Verfügung stehen. Da dieses Phänomen, wie bereits erwähnt, verstärkt bei sprachschwächeren bzw. mit bestimmten Textsorten und -mustern weniger vertrauten Schülerinnen und Schülern, die bestimmte Routinen noch nicht internalisiert haben, auftreten sollte, sollten inhaltlich-stilistisch weniger leistungsstarke Schülerinnen und Schüler auch verstärkt im orthografisch-grammatischen Bereich niedrigere Leistungen zeigen und vice versa.

[36] Auch die Dimension s*prachliche Richtigkeit* erweist sich am Ende der Sekundarstufe I als höher mit *Inhalt* und *Stil* assoziiert als am Ende der Primarstufe (Inhalt: $.48_{Primar}$ vs. $.60\text{-}.69_{SekI}$; Stil: $.65_{Primar}$ vs. $.71\text{-}86_{SekI}$; Böhme et al., 2009, S. 319).

Auch die Unterschiede zwischen den beiden anderen Textmustern (*argumentieren* und *narrativ*), das heißt den höheren Zusammenhang zwischen *Inhalt* und *sprachliche Richtigkeit* für das Argumentieren, können sich unter Annahmen der geteilten kognitiven Ressourcen erklären lassen. So ist das Erzählen eine (auch bereits für Kinder) alltagsrelevante Kommunikationsform, deren aufgabenübergreifende inhaltliche Aspekte (wie etwa das Arrangement der Inhalte) bereits früh eingeübt werden (Ehlich, 1980; Quasthoff, 1980). Darüber hinaus stellen narrative Aufgaben, vor allem diejenigen, die einer klassischen Erzählstruktur folgen, häufig im inhaltlichen Bereich nur wenige Anforderungen bereit, welche über Basisanforderungen hinausgehen. Oftmals genügt das korrekte Aufgreifen des Themas und eine Realisierung und lineare Ordnung der Erzählphasen *Exposition*, *Komplikation*, *Höhe-* oder *Wendepunkt* und *Auflösung*. Für das Argumentieren liegen deutlich mehr inhaltliche Anforderungen vor, so etwa die hinreichende Stützung und logische Struktur eines jeden Arguments, eine hinreichende Menge der Argumente oder bei dialektischen Argumentationen auch das Vorhandensein von Argumenten der beiden kontrastierten Positionen. Aufgrund der vergleichsweise niedrigen inhaltlichen Anforderungen für die Bearbeitung narrativer Aufgaben kann davon ausgegangen werden, dass die inhaltliche Auseinandersetzung mit dem Thema nicht so viele zusätzliche kognitive Ressourcen bindet, sodass andere, d. h. orthografisch-grammatische Schreibteilleistungen dadurch in Mitleidenschaft gezogen würden (Kellogg, 1987). Dies steht auch im Einklang mit dem in Kapitel 4 berichteten Befund, dass in der Argumentation geübtere Schülerinnen und Schüler auch im orthografisch-grammatischen Bereich höhere Leistungen bei der Bearbeitung von Aufgaben dieses Textmusters erzielen (vgl. 4.3.2.3 und 4.4.).

Generell sind diese feineren Unterschiede hinsichtlich der internen Struktur im Vergleich der Textmuster jedoch vorsichtig zu interpretieren; die Datenbasis pro Textmuster umfasst jeweils nur Leistungsdaten von vier verschiedenen Aufgaben. Ausgewogenheiten (beispielsweise an Textsorten, thematischer Variabilität oder Prototypizitätsgraden) können in dieser Größenordnung kaum gewährleistet werden. Darüber hinaus besteht jeweils nur eine direkte Verbindung (Aufgabenverknüpfung) zwischen je zwei Textmustern. Hier müssen weitere Untersuchungen unter Einsatz mehrerer Textmusterverbindungen und unter Kontrolle thematischer Aspekte und Textsorteneigenschaften zeigen, inwiefern im Rahmen dieser Studie gefundene Unterschiede möglicherweise auf solche Aspekte zurückzuführen sind.

Zusammenfassend legen die Evidenzen aus dieser Studie nahe, dass Schreibkompetenzen von Schülerinnen und Schülern am Ende der Sekundarstufe I textmusterspezifisch sind. Diese

Kompetenzen unterliegen textmusterunabhängig einer zweidimensionalen inneren Struktur mit den Faktoren *Inhalt+Stil* einerseits und *sprachliche Richtigkeit* andererseits. Die Dimension der sprachlichen Richtigkeit scheint hierbei ein textmusterunabhängiges Teilkonstrukt, während inhaltliche und stilistische Schreibkompetenzen textmusterspezifisch sind.

7. Miterfassung von Lesekompetenz bei der Messung von Schreibkompetenz (Teilstudie II)

Ziel der in diesem Kapitel vorgestellten Teilstudie ist es, zu untersuchen, ob und in welchem Umfang empirisch ermittelte Schreibkompetenzen dadurch verzerrt sind, dass bei der Messung auch Lesefähigkeiten miterhoben werden. Zur Beantwortung dieser Frage werden Zusammenhänge zwischen Lese- und Schreibfähigkeiten in Abhängigkeit von leseschwierigkeitsbestimmenden Merkmalen der verwendeten Schreibaufgaben untersucht.

7.1. Schreibkompetenzmessung und Rezeptionsfähigkeiten

Wie in Kapitel 5 dargelegt, ist eine validitätsbedrohende oder zumindest validitätsmindernde Gefahrenquelle, was Messick (1990, 1996) mit *construct-irrelevant variance* beschreibt, d. h. die Möglichkeit, dass im Rahmen der Testung eines Konstrukts konstruktexterne Aspekte miterfasst werden, also Aspekte, welche einem anderen Konstrukt zuzuordnen sind. Als Gebot formuliert, spricht Messick von *minimal construct-irrelevant variance*. Für die Messung von Schreibkompetenz impliziert dies eine Erfassung des Konstrukts *Schreibkompetenz* unter (weitestmöglichem) Ausschluss von anderen nichtsprachlichen und sprachlichen Fähigkeiten, so auch unter Ausschluss von Lesefähigkeiten, insofern diese nicht auch Teil des Konstrukts *Schreibkompetenz* sind.

Um Schreibkompetenzen jedoch standardisiert messbar zu machen, müssen den Getesteten Aufgaben vorgelegt werden, die in einem hinreichenden Maße Vorgaben machen, damit das Geschriebene im Anschluss hinreichend präzise und reliabel ausgewertet werden kann. Bestimmte stilistische und inhaltliche Anforderungen müssen aus der Aufgabenstellung hervorgehen, damit mit entsprechenden Skalen und Auswertungskriterien (vgl. Kapitel 3) überprüft werden kann, ob bzw. inwieweit diese Anforderungen erfüllt werden. Somit setzt die Bearbeitung einer Schreibaufgabe die erfolgreiche Rezeption einer mehr oder weniger komplexen Instruktion und ergänzendem Material in Form eines zugrunde liegenden Stimulus voraus (vgl. die Beispielaufgaben im Anhang unter A.3.1.1 und A.3.1.2).

Des Weiteren stellen inhaltliche und stilistische Vorgaben in den Aufgaben eine Engführung der Aufgabenstellung dar und grenzen die Informationsmenge ein, womit den Getesteten ein

klarer und eindeutiger Arbeitsauftrag vermittelt wird. Gerade in einer Testsituation und einem diesem Test zugrunde liegenden begrenzten Zeitrahmen ist es wichtig, den Getesteten die Anforderungen der Aufgabe transparent zu machen. Darüber hinaus dient die Vorgabe inhaltlich relevanter Informationen dazu, Schreibkompetenzen weitgehend unabhängig von dem Weltwissen der Getesteten zu prüfen. Lägen keine inhaltlichen Vorgaben vor, müssten die Getesteten auf ihren Erfahrungs- und Wissensschatz zurückgreifen, womit die Leistungen in erheblichem Maße vom Vorwissen der Testpersonen beeinflusst wären, die Ergebnisse durch Wissenskomponenten verzerrt würden.

Die Praktikabilität einer objektiven und reliablen Schreibkompetenzmessung und das Gebot der minimalen konstruktexternen Varianz stehen somit in einem gewissen Widerspruch zueinander. Eine vollkommen von Rezeptionskompetenzen losgelöste, aber dennoch systematische Messung von Schreibkompetenzen ist kaum möglich, eine Miterfassung von Rezeptionskompetenzen bzw. mindestens einer Rezeptionskompetenz somit unumgehbar. Eine bislang ungeklärte Frage ist jedoch, in welchem Umfang die für die erfolgreiche Aufgabenbearbeitung vorausgesetzte Rezeptionskompetenz die Leistungen im Test zur Messung der Zielkompetenz beeinflussen.

7.2. Vier Basissprachkompetenzen und ihre Zusammenhänge

Sprachliche Fähigkeiten lassen sich in vielfältiger Weise kategorisieren. Eine vorherrschende basale Kategorisierung sprachlicher Fähigkeiten ist die nach Medium und Verarbeitungsweg (Harris, 1969; Pfister & Kaufmann, 2008; Tatham & Morton, 2006; Treiman, Clifton, Meyer & Wurm, 2003). Der Ausdruck *Verarbeitungsweg* bezieht sich darauf, ob Sprache rezipiert oder produziert wird. Unter *Medium* (häufig auch *Modus*) versteht man die Art und Weise, das Vehikel, wie Sprache kommuniziert, transportiert, übertragen wird. Darunter fallen die gesprochene Sprache, die (optische) Schriftsprache, die Blindenschrift oder die Gebärdensprache. Für sinnlich unbeeinträchtigte Mitglieder unseres Kulturkreises sind hierbei die gesprochene und die (optisch) geschriebene Sprache zentral und für nahezu jedermann relevant für eine erfolgreiche gesellschaftliche Teilhabe. Kreuzt man nun diese beiden relevanten Modi mit den Verarbeitungswegen, gelangt man zu den in Tabelle 7.2.1. angeführten vier sprachlichen Basiskompetenzen, welche auch der Standardklassifikation von sprachlichen Fertigkeiten in linguistisch-didaktischen Taxonomien entsprechen (Jude, 2008; Messelken 1971; Schöler, 2006):

Tabelle 7.2.1: ***Sprachliche Basiskompetenzen***

	Rezeption	**Produktion**
gesprochen	Zuhören	Sprechen
geschrieben	Lesen	Schreiben

Diese vier sprachlichen Basiskompetenzen sind jedoch in weiten Teilen nicht unabhängig voneinander, da ihnen teilweise identische kognitive Prozesse zugrunde liegen. Psycholinguistische Evidenzen und Theorien belegen, dass beim Sprachverstehen und bei der Sprachproduktion teilweise mediumsunabhängige, teilweise mediumsabhängige Prozesse beteiligt sind. Im Bereich der Produktion geht man zunächst von einer abstrakten, d. h. formfreien *Message*-Bildung aus, die dann in weiteren Schritten formuliert und schließlich artikuliert wird. Erst im Endstadium der Formulierung wird auf phonologische und graphematische und somit mediumsspezifische sprachliche Merkmale zugegriffen (Kempen & Hoenkamp, 1987; Levelt, 1989). Auch für die Rezeption erweisen sich zahlreiche Mechanismen wie etwa das syntaktische Dekodieren (*Parsing*) oder der Aufbau von propositionalen Netzwerken als weitgehend mediumsunabhängig (Danks & End, 1987; Gernsbacher, Varner & Faust, 1990; Kürschner & Schnotz, 2008; Maia, 2008; Pollatsek, Ashby & Clifton, 2012; Weidenmann 1997), allerdings unterscheiden sich das Lesen und das Zuhören in basalen Wahrnehmungseigenschaften (Aufnahme eines optischen vs. akustischen Reizes), im Verarbeitungstempo sowie im Rahmen der Verarbeitung größerer sprachlicher Einheiten (Texte), in den Anforderungen an das Arbeitsgedächtnis und im gegebenen Orientierungsrahmen (Baddeley, 1997; Baddeley, Thomson & Buchanan, 1975; R. A. Cohen, Sparling-Cohen & O'Donnell 1993; Cutler & Clifton 1999; Imhof 2003; Kürschner & Schnotz, 2008; O. Neumann, van der Heijden & Allport, 1986; Perfetti 1999). Ein weiterer Zusammenhang zwischen *Lesen* und *Zuhören* liegt auch darin begründet, dass beim Lesen das Gelesene implizit verlautlicht wird. Coltheart (1978) beschreibt dies im *Dual-Route-Modell*, nach welchem die Verarbeitung schriftlicher Information auf zwei Wegen erfolgt, zum einen über einen spezifischen schriftsprachlichen Rezeptionsweg, zum anderen über phonologisches Rekodieren. Dabei können die Stärken der beiden Verarbeitungswege unterschiedlich stark ausgeprägt sein; mit ansteigender Lesefrequenz und Übung im Lesen steigt auch die Stärke des schriftsprachspezifischen Verarbeitungsweges (Niznikiewicz & Squires, 1996).

Im Vergleich der Sprachproduktion und der Sprachrezeption erweisen sich die beiden Vorgänge keineswegs als reziprok. So verläuft der Rezeptionsprozess in weiten Teilen parallel, viele Informationen werden gleichzeitig verarbeitet (Rickheit et al., 2003; Rickheit & Strohner, 1993), die Sprachproduktion läuft in weiten Teilen seriell bzw. inkrementell ab (Kempen & Hoenkamp, 1987; Levelt, 1989). Gemeinsam ist den Verarbeitungsprozessen der Zugriff auf Wissensbestände wie etwa den Wortschatz oder das Phonem- und Grapheminventar im mentalen Lexikon (Aitchison, 1997; Engelkamp, 1995; Grohnfeldt, 2007). Auch wenn die Wissensstrukturen selbst unabhängig vom Verarbeitungsweg sind, so ist die Art (und auch die damit verbundene Leichtigkeit/Schwierigkeit) des Zugriffs verarbeitungswegspezifisch. So ist der rezeptive (passive) Wortschatz deutlich größer als der produktive (aktive) Wortschatz (Klann-Delius, 2008; Steinhoff, 2009). Dies lässt sich auf allgemeine Erinnerungsprozesse zurückführen, dass das Wiederkennen viel schneller und besser gelingt als das Reproduzieren (u. a. Loftus, 1971; Tversky, 1973).

Im Einklang mit den psycholinguistischen Evidenzen konnten im Rahmen bisheriger Leistungsstudien, in welchen mehrere der sprachlichen Basiskompetenzen untersucht wurden, mittlere bis hohe Zusammenhänge festgestellt werden. Für die rezeptiven Fähigkeiten *Lesen* und *Zuhören* berichtet Jude (2008, S. 43) in der Betrachtung mehrerer Studien über ein Korrelationsspektrum zwischen $r = .45$ und $r = .75$ (vgl. auch Diakidoy et al., 2005; Rost & Hartmann, 1992). Bremerich-Vos, Böhme und Robitzsch (2009, S. 210) berichten für den Primarbereich latente Korrelationen zwischen *Lesen* und *Zuhören* von $r = .85$. Ebenso konnten die Autoren hohe Korrelationen zwischen den schriftsprachlichen Kompetenzen *Lesen* und *Schreiben* ermitteln ($r = .70$), jedoch nur einen Zusammenhang mittlerer Stärke zwischen *Schreiben* und *Zuhören* ($r = .44$). Einen hohen Zusammenhang ($r = .80$, S. 1) zwischen der Entwicklung von Schreib- und Lesekompetenz konnte auch Smith (2009), untermauert von Ausführungen von Prose (2006), nachweisen. Weitere Evidenzen finden sich bei Heller (1999) oder Parodi (2007).

Für den Bereich *Sprechen* liegen bisher nur wenige systematisch erhobene Daten für die Erstsprache vor und bislang keine, in welchen Sprechkompetenzen und weitere sprachliche Kompetenzen parallel erhoben wurden. So untersuchten Berninger et al. (2006) den Zusammenhang zwischen *Sprechen* und den anderen sprachlichen Fähigkeiten *Lesen*, *Schreiben* und *Zuhören*, für das Sprechen wurde jedoch lediglich ein Teilaspekt erfasst, *oral expression* (≈ Ausdrucksfähigkeit). Allerdings zeigen Untersuchungen aus dem Zweitsprach-

erwerb[37], dass mittlere bis hohe Zusammenhänge sowohl zwischen gesprochen-sprachlichen Fähigkeiten *Sprechen* und *Zuhören* (gesprochene Sprache) (Liao, Qu & Morgan, 2010), als auch zwischen den produktiven Fähigkeiten *Sprechen* und *Schreiben* (ETS, 2010; Hubert, 2008; Hubert, 2013; Powers, Kim, Yu, Weng & VanWinkle, 2009) bestehen. In der Untersuchung von Liao und Kollegen wurden alle vier sprachlichen Basisfähigkeiten (*Zuhören*, *Lesen*, *Sprechen* und *Schreiben*) erfasst; es zeigten sich durchweg bedeutsame Zusammenhänge, am schwächsten ausgeprägt waren jedoch die Zusammenhänge für *Schreiben* und *Zuhören* sowie für *Lesen* und *Sprechen* (S. 13.4).

Aus den Ergebnissen dieser Studien bleibt festzuhalten, dass …

(i) mittlere bis hohe Zusammenhänge zwischen allen vier Basiskompetenzen bestehen, dass jedoch höhere Zusammenhänge zwischen Kompetenzen vorliegen, insofern die beiden Kompetenzen denselben Verarbeitungsweg betreffen oder im selben sprachlichen Medium stattfinden;
(ii) diese Zusammenhänge in der Sprachentwicklung zunächst zunehmen, ab einem hinreichenden Maß der Sprachexpertise jedoch rückläufig sind bzw. sein können.

(i) legt nahe, dass es offenbar neben spezifischen Sprachkompetenzanteilen je gemeinsame Rezeptions- (*Zuhören* und *Lesen*), Produktions- (*Sprechen* und *Schreiben*) und mediums-spezifische (jeweils *Zuhören* und *Sprechen* sowie *Lesen* und *Schreiben*) Kompetenzanteile sowie einen sprachallgemeinen Kompetenzanteil gibt.

Die in Tabelle 7.2.1 angeführten vier sprachlichen Basiskompetenzen *Lesen*, *Zuhören*, *Schreiben* und *Sprechen* sind somit empirisch nicht unabhängig. Aufgrund der teilweise geteilten, teilweise bereichsspezifischen kognitiven Prozesse und Ressourcen kann davon ausgegangen werden, dass es kompetenzspezifische, verarbeitungswegsspezifische, mediumsspezifische und von diesen Differenzierungen unabhängige sprachallgemeine Fähigkeiten gibt, welche den Basiskompetenzen zugrunde liegen. Abbildung 7.2.1 illustriert eine solche Klassifikation dieser Fähigkeiten.

[37] In der Spracherwerbsforschung war die Identität vs. Differenz von Erst- und Zweitspracherwerb lange Zeit ein strittiges Thema (Bley-Vroman, 1989; Carroll, 2001; Ervin-Tripp, 1974; Ferguson, 1962; Flynn, 1996; Lado, 1957; Meisel, Clahsen & Pienemann, 1981; Schwartz & Sprouse, 1996; L. White, 1989). Inzwischen geht man davon aus, dass die Differenz bzw. Identität der im Erst- und Zweitspracherwerb beteiligten Prozesse vom Lebensabschnitt und der Art und Weise des Erwerbs abhängen (Cummins, 2006; Klein, 2007; Klein & Dimroth, 2009). Auch wenn Evidenzen aus der Zweitspracherwerbsforschung somit nicht unmittelbar auf erstsprach-orientierte Situationen übertragbar sind, können diese dennoch als Indizien für mögliche analoge Prozesse und Strukturen herangezogen werden.

Abbildung 7.2.1: ***Neunfelderschema: Sprachliche Fähigkeiten.***

Spezifische Zuhörfähigkeiten	**Allgemeine Sprechsprachfähigkeiten**	**Spezifische Sprechfähigkeiten**
Allgemeine Sprachrezeptionsfähigkeiten	**Allgemeine Sprachkompetenz**	**Allgemeine Sprachproduktionsfähigkeiten**
Spezifische Lesefähigkeiten	**Allgemeine Schriftsprachfähigkeiten**	**Spezifische Schreibfähigkeiten**

weiß: sprachbereichsspezifische Fähigkeiten; hellgrau: mediums- oder verarbeitungswegsspezifische Fähigkeiten; dunkelgrau: bereichsübergreifende Sprachfähigkeiten

7.3. Präzisierung der Fragestellung

Zurückkehrend zum Ausgangspunkt dem theoretisch idealen Ziel, *Schreibkompetenz* so unabhängig wie möglich von Rezeptions- bzw. im konkreten Fall *Lesekompetenz* zu messen, bleibt festzuhalten, dass die Überprüfung, zu welchem Grad diesem Anspruch gerecht geworden werden kann, nicht einfach durch bloße Ermittlung von Zusammenhängen zwischen diesen Kompetenzen erfolgen kann. Wie dargelegt, gibt es allgemeinsprachliche Fähigkeiten, die allen Rezeptions- und Produktionskompetenzen zugrunde liegen sowie mediumsspezifische Fähigkeiten, wobei allgemeine schriftsprachliche Fähigkeiten sowohl Teil der Lesekompetenz als auch Teil der Schreibkompetenz sind. Der Schreibkompetenz liegen somit im Rahmen des in Abbildung 7.2.1. veranschaulichten Schemas die folgenden vier Komponenten zugrunde: allgemeine Sprachfähigkeiten, welche auch jeweils den anderen drei Basiskompetenzen zugrunde liegen, allgemeine Sprachproduktionsfähigkeiten, welche auch einen relevanten Teil der Sprechkompetenz darstellen, allgemeine Schriftsprachfähigkeiten, welche auch der Lesekompetenz zugrunde liegen, sowie spezifische Schreibfähigkeiten, die unabhängig von den anderen drei Basiskompetenzen sind (vgl. Abbildung 7.3.1).

Stellt man der Schreibkompetenz die Lesekompetenz gegenüber, verdeutlicht dies die gemeinsamen Anteile (vgl. Abbildung 7.3.2.).

Abbildung 7.3.1: ***Schreibkompetenz: geteilte und ungeteilte Kompetenzanteile.***

Allgemeine Sprachfähigkeiten	**Allgemeine Sprachproduktionsfähigkeiten**
Allgemeine Schriftsprachfähigkeiten	**Spezifische Schreibfähigkeiten**

= Schreibkompetenz

weiß: von Lesekompetenz unabhängige Schreibkompetenzanteile; grau: gemeinsame Kompetenzanteile von *Lesen* und *Schreiben*.

Abbildung 7.3.2: ***Lesekompetenz: geteilte und ungeteilte Kompetenzanteile.***

Allgemeine Sprachrezeptionsfähigkeiten	**Allgemeine Sprachfähigkeiten**
Spezifische Lesefähigkeiten	**Allgemeine Schriftsprachfähigkeiten**

= Lesekompetenz

weiß: von Schreibkompetenz unabhängige Lesekompetenzanteile; grau: gemeinsame Kompetenzanteile von *Lesen* und *Schreiben*.

Somit präzisiert sich die Ausgangsfragestellung wie folgt: Inwiefern werden bei der Messung von Schreibkompetenzen mit einer schriftsprachlichen Präsentation der Aufgabenstellung[38] spezifische Lesekompetenzanteile miterfasst? Stellt die schriftliche Rahmung der Aufgabe einen so hohen Anspruch an die Lesekompetenz, dass die erfolgreiche Bearbeitung der Schreibaufgaben von spezifischen Lesefähigkeiten abhängt? Zeigen sich Zusammenhänge zwischen Lese- und Schreibkompetenz, welche über den auf geteilten Kompetenzanteilen beruhenden Basiszusammenhang hinausgehen?

Da es sich bei dem Basiszusammenhang um unkritische gemeinsame Kompetenzanteile handelt, da diese auf identischen Sprachverarbeitungsprozessen beruhen, muss ein Maß gefunden werden, um lesespezifische Kompetenzanteile zu quantifizieren. Die Fragestellung lässt sich somit idealiter auf der Basis eines Sets an Schreibaufgaben untersuchen, die unterschiedliche Anforderungen an die spezifische Lesekompetenz stellen. Insofern sich diese Anforderungen kriterial erfassen lassen, kann die allgemeine Fragestellung konkretisiert werden: Zeigen sich höhere bzw. niedrigere Zusammenhänge zwischen den Kompetenzen der

[38] Da in der zugrunde liegenden Studie die Schreibaufgaben stets literarisch-visuell dargeboten wurden, d. h. in Form von Text und ggf. Bildern, und akustische Instruktionen nicht zum Einsatz kamen, kann in diesem Rahmen lediglich der Einfluss der Lesekompetenz (als relevanter Rezeptionskompetenz) untersucht werden.

Bereiche *Lesen* und *Schreiben* in Abhängigkeit von leseschwierigkeitsbestimmenden Merkmalen der Stimuli der zugrunde liegenden Schreibaufgaben?

7.4. Schwierigkeitsbestimmende Aufgabenmerkmale

Textuelle Aufgaben und Aufgabenteile variieren in ihrer Schwierigkeit auf verschiedenen Ebenen und können diesbezüglich kategorisiert werden (Christmann & Groeben, 1999; Groeben, 1982; Hochhaus, 2004; Rosebrock, 2012):

a) sprachliche Einfachheit
b) kognitive Gliederung
c) inhaltliche Aspekte
d) motivationale Stimulanz

Sowohl für das (mediumsunabhängige) Instruktionsverstehen als auch für das (schriftsprachspezifische) Textverstehen wurden zahlreiche schwierigkeitsbestimmende Merkmale herausgearbeitet (Draxler, 2005; Freedle & Kostin, 1993; Kauertz, 2007; Köhler & Altmann, 1986; Köster, 2005; Nold & Rossa, 2007; Nunan & Koebke, 1995; Prabhu, 1987; Schuman & Eberle, 2011; Schweitzer 2007; Willenberg, 2007). Im Rahmen der hier vorgestellten Teilstudie wird auf Aufgabenmerkmale, die Aspekte der sprachlichen Einfachheit erfassen, fokussiert. Bei den erfassten Aspekten handelt es sich um:

- Textmenge / Textlänge
- sprachliche Komplexität
- lexikalisches Niveau

Die Beschränkung auf diese Merkmale basiert auf mehreren Gründen: Zum einen lassen sich diese Merkmale objektiv (bspw. durch die Zählung von Wörtern oder Sätzen) erfassen, bei den verwendeten Kriterien handelt es sich um bereits etablierte. Einstufungen etwa von *inhaltlicher Komplexität*, erfordern weitere Festlegungen oder Einschätzungen von Experten und ggf. weitere Überprüfungen (z. B. Reliabilitätsmessungen) dieser Einschätzungen (Chalifour & Powers, 1989). Dies ist mit einem enormen zusätzlichen Zeit-, Personen- und Kostenaufwand verbunden. Außerdem liegen bisher auch nur wenige in Einzelstudien verwendete Kategorisierungs- und Analysesysteme vor. Zum anderen sind die heran-gezogenen Aufgaben gezielt für die Testung in einem einheitlichen zeitlichen Rahmen (20

Minuten Testzeit) für eine bestimmte Gruppe an Testpersonen (Schülerinnen und Schüler der 8. bis 10. Jahrgangsstufe des allgemeinbildenden Schulsystems der Bundesrepublik Deutschland) und im Hinblick auf eine vergleichbare Auswertung mit hinreichend ähnlichen Schemata (vgl. Kapitel 3) entwickelt worden. Daher weisen die Aufgaben in Aspekten, welche etwa der inhaltlichen Komplexität zuzuordnen wären, bereits konzeptionell kaum Varianz auf.[39]

Darüber hinaus ist die Einschränkung auf Aspekte der sprachlichen Einfachheit auch konzeptuell motiviert. Die Heranziehung von leseschwierigkeitsbestimmenden Aspekten ist nur insofern zielführend im Rahmen der vorliegenden Fragestellung, als dass diese Aspekte ausschließlich Aspekte betreffen, welche für das Lesen relevant sind, nicht jedoch für das Schreiben. Da das Aufgreifen und Anführen von inhaltlichen Aspekten explizit bei der Bewertung der Schreibleistungen berücksichtig wird (vgl. Kapitel 3), erweisen sich diese Aspekte als ungeeignet, um konstruktirrelevante Varianz zu quantifizieren. Sprachliche Merkmale, welche die Rezeptionsschwierigkeit bestimmen, sind hingegen für das freie Schreiben irrelevante Merkmale, welche somit zumindest einen Teil des schreibkonstrukt-irrelevanten Anteils der Lesekompetenz quantifizieren können.[40]

Konkret wurden zur Bestimmung der sprachlichen Anforderungen an die Lesekompetenz die im Folgenden erläuterten Merkmale bestimmt.

7.4.1. Textlänge

Zur Bestimmung der Textlänge dienten vier Maße, welche auf unterschiedlichen sprachlichen Ebenen ansetzen und die Textlänge anhand der entsprechenden Basiseinheiten bestimmen:

- Anzahl der Zeichen: Als Zeichen wurden Buchstaben, Ziffern und Satzzeichen gezählt, Leerzeichen wurden nicht berücksichtigt.
- Anzahl der Silben

[39] Dies spiegelt sich beispielsweise auch darin wider, dass sich für alle Aufgaben die inhaltlichen Anforderungen in Form von 4 bis 5 Kriterien erfassen ließen (vgl. Kapitel 3).

[40] Auch ein Vergleich mit den Bildungsstandards, welche die konzeptionelle Grundlage für das hier vorliegende Konstrukt *Schreibkompetenz* bilden, findet sich lediglich unter dem speziellen Unterpunkt *Ergebnisse einer Textuntersuchung darstellen* die Anforderung, auch längere und komplexere Texte rezipieren zu können. Dies betrifft jedoch den Spezialfall des textbasierten Schreibens und nicht das freie Schreiben, das Grundlage der Normierungsstudie, der vorgestellten Kompetenzstufenmodelle und der hier vorgestellten Forschungsstudien ist.

- Anzahl der Wörter: Wörter wurden nach dem orthografischen Kriterium erfasst; als Wörter zählten alle Zeichenketten, die zwischen zwei Leerzeichen (oder Analoga wie Satzzeichen oder Absatzgrenzen) stehen (Kessel & Reimann, 2012).
- Anzahl der Sätze: Als Sätze wurden alle Abfolgen sprachlicher Einheiten (in der Regel Wörter) gezählt, welche mit einem Satzschlusszeichen enden.

Motiviert ist die Berücksichtigung dieser verschiedenen Kategorisierungsebenen dadurch, dass sie an verschiedenen Teilprozessen des komplexen Prozesses des Leseverstehens ansetzen: Beim Leseverstehen finden unterschiedliche Operationen statt; so muss der Lesende auf der Wortebene Buchstaben zu einem Wort zusammenfügen und dieses als Wort erkennen (lexikalische und morphologische Re- und Dekodierung) (Perfetti, 1992; Stanovich, 2000; Verhoeven & Perfetti, 2011; Verhoeven & Carlisle, 2006; Verhoeven & van Leeuwe, 2009). Auf Satzebene müssen die Wortabfolgen wiederum durch syntaktische Dekodierung (*Parsing*) und semantische Analyse in Satzbedeutungen (= eine oder mehrere Propositionen) überführt werden (Frederiksen, 1975; Rayner, Carlson & Frazier, 1983). Auf Textebene müssen die Sätze beziehungsweise deren Bedeutungen zueinander in Beziehung gesetzt, Inferenzen gebildet, ein mentales Modell aufgebaut werden (Johnson-Laird, 1983; Kintsch & van Dijk, 1978; McKoon & Ratcliff, 1992; Singer, 1990).

Die Anzahl der Silben als kleinere Einheiten von Wörtern wurde parallel zur Anzahl der Zeichen erfasst, um einerseits die Unterschiede der beiden Verarbeitungswege beim Lesen (spezifisch schriftsprachlich vs. auditiv aufgrund impliziter Verlautlichung)[41] zu erfassen, andererseits um die Basiseinheiten für komplexere (Kombinations-)Maße zur Erfassung der sprachlichen Schwierigkeit zu bestimmen (vgl. folgende Ausführung unter 7.4.4.).

7.4.2. sprachliche Komplexität

Auch die sprachliche Komplexität wurde mit mehreren Maßen erfasst, die wiederum den unterschiedlichen Lese(teil)prozessen Rechnung tragen:

- Zeichen pro Wort (= graphematische Komplexität der Wörter)
- Silben pro Wort (= phonologische Komplexität der Wörter)
- Wörter pro Satz (= syntaktische Komplexität der Sätze)

[41] Vgl. die Ausführungen in Kapitel 7.2. zum *Dual-Route-Modell.*

7.4.3. lexikalisches Niveau

Um das lexikalische Niveau der Stimulus- und Instruktionstexte zu bestimmen, wurde für alle in diesen Texten vorkommenden Wörter die jeweilige Häufigkeitsklasse bestimmt. „Mit Häufigkeitsklassen werden alle Wörter des gesamten Vokabulars [einer Sprache] nach ihrer Häufigkeit in Klassen aufgeteilt, wobei Wörter derselben Klasse ungefähr gleich häufig sind“ (Keibel, 2008, 2009). Die Häufigkeitsklasse bestimmt sich hierbei als Logarithmus zur Basis 2 des Quotienten [Ziel-Wort] ÷ [häufigstes Wort im Vokabular]. Insgesamt werden die Wörter in 30 Klassen unterschieden, welche aufsteigend ganzzahlig durchnummeriert sind; die Häufigkeitsklasse 0 enthält das häufigste Wort des Vokabulars. Die Bestimmung der Häufigkeitsklassen der Wörter erfolgte über das Wortschatzportal der Universität Leipzig („Wortschatz Universität Leipzig“, 2014).

Für jede Schreibaufgabe wurden nun drei Maße der lexikalischen Schwierigkeit des Stimulus- und Instruktionstextes bestimmt:

- durchschnittliche Häufigkeit der enthaltenen Wörter (= Mittelwert der Häufigkeitsklassenwerte aller enthaltenen Wörter)
- Anzahl seltener Wörter (= absolute Anzahl der Wörter mit einer Häufigkeitsklasse > 15)[42]
- Anteil seltener Wörter (= relative Anzahl der Wörter mit einer Häufigkeitsklasse > 15, = Anzahl seltener Wörter ÷ Anzahl aller vorkommender Wörter)

Die Erfassung der durchschnittlichen Häufigkeit aller Wörter trägt jedem Seltenheitsunterschied im Vokabular Rechnung; allerdings ist fraglich, ob die Unterschiede im häufigen und mittleren Frequenzbereich auch für den Sprachnutzer einen Unterschied darstellen. In der Lesbarkeitsforschung werden üblicherweise die Anzahl und/oder der Anteil der seltenen Wörter als schwierigkeitsbestimmendes Merkmal dargestellt (Bamberger & Vanecek, 1984; DuBay, 2007).

[42] Es finden sich in der Literatur keine genauen Grenzfestlegungen, ab wann ein Wort als selten gilt, jedoch werden Beispiele für das mittlere Frequenzspektrum aus den Häufigkeitsklassen 10-12 angeführt, sehr seltene für einen Klassenbereich um 20 (u. a. Keibel, 2008, 2009). Der Grenzwert für die vorliegende Kategorisierung wurde aufgrund dieser exemplarischen Angaben auf 15 gesetzt.

7.4.4. Kombinationsmaße zur Erfassung der Lesbarkeit von Texten

Zusätzlich wurden zwei etablierte Lesbarkeitsmaße eingesetzt, der LIX (Lesbarkeitsindex) sowie der Flesch-Index (*Flesch reading ease*).[43]

Der LIX geht auf Björnsson (1968) zurück und bestimmt sich als Summe der durchschnittlichen Satzlänge (gemessen in Wörter pro Satz) und dem prozentualen Anteil an langen Wörtern (= Wörter mit mehr als sechs Zeichen). Der LIX steigt mit zunehmender Schwierigkeit an, als Richtwerte dienen: < 40 = leicht; > 60 = schwer.

Gemäß dem von Rudolf Flesch für das Englische entwickelten und von Toni Amstad auf das Deutsche angepassten Flesch-Index werden folgende Parameter zur Berechnung herangezogen: durchschnittliche Satzlänge (gemessen in Wörter pro Satz) und durchschnittliche Wortlänge in Silben. Der Index bestimmt sich schließlich wie folgt: 180 − Ø Satzlänge − (58.5 × Ø Silben pro Wort). Der Wertebereich bewegt sich zwischen 100 (sehr leicht) und 0 (sehr schwer), als Richtwerte dienen: > 70 = leicht, < 20 = schwer (Flesch, 1948; Amstad 1978).

Beide Maße beziehen die syntaktische Komplexität (Wörter pro Satz) und indirekt das lexikalische Niveau mit ein. Die Verwendung von Längenmaßen (lange Wörter bzw. Anzahl der Silben) beruht auf Evidenzen der Quantitativen Linguistik, dass die Wortlänge mit der Worthäufigkeit hoch assoziiert ist; häufigere Wörter sind kürzer, seltenere länger (Best, 2005, 2006b; G. Wimmer & Altmann, 1996; G. Wimmer, Köhler, Grotjahn & Altmann, 1994). Ein wesentlicher Unterschied zwischen dem LIX und dem Flesch-Index ist jedoch, dass für den LIX nur die seltenen Wörter (oberhalb einer gesetzten Seltenheitsgrenze) betrachtet werden, für den Flesch-Index alle Wörter miteinbezogen werden (vgl. die Ausführungen in 7.4.3.). Ein weiterer Unterschied ist, dass der LIX sich an der schriftsprachlichen Form der Wörter (Anzahl der Zeichen), der Flesch-Index an der phonologischen Struktur (Anzahl der Silben) orientiert.

[43] Die Auswahl genau dieser beiden Lesbarkeitsindizes war dadurch motiviert, dass die Aufgabenentwicklerinnen und -entwickler (vgl. Kapitel 3.1.), welche auch die dieser Studie zugrunde liegenden Leseaufgaben generierten, mit diesen beiden Kennwerten im Rahmen der Textauswahl operierten.

7.5. Hypothesen

Aufgrund der teilweise gemeinsamen kognitiven Sprachverarbeitungsprozesse beim Lesen und beim Schreiben (vgl. Kapitel 7.2.) wird angenommen, dass sich auch in der vorliegenden Studie ein moderater bis hoher Basiszusammenhang zwischen Schreib- und Lesefähigkeiten festgestellt werden kann.

Des Weiteren wird erwartet, dass der Zusammenhang zwischen Lese- und Schreibkompetenz von den schreibaufgabenspezifischen Anforderungen an die Lesekompetenz abhängt und der Zusammenhang mit zunehmender sprachlicher, vor allem grammatischer Komplexität, mit zunehmender Textmenge und mit zunehmendem lexikalischem Niveau steigt.

7.6. Durchführung der Studie

7.6.1. Datengrundlage

Die Datengrundlage lieferte eine Studie zur Schreibkompetenz- und Lesekompetenzerfassung aus dem Jahr 2010, welche auch zur Pilotierung einiger der Normierungsaufgaben diente (vgl. Kapitel 3.2.). An der Studie nahmen insgesamt 1726 Schülerinnen und Schüler der achten Jahrgangsstufe aller Schulformen des allgemeinbildenden Schulsystems der Bundesrepublik Deutschland teil. Der Altersdurchschnitt betrug 14.7 Jahre (*SD* = 0.66 Jahre), 46 % der Stichprobe waren weiblich.

Im Rahmen dieser Studie wurden von 80 % der teilnehmenden Schülerinnen und Schüler jeweils mehrere Leseaufgaben und zumeist zwei Schreibaufgaben bearbeitet. Insgesamt wurden 19 Leseaufgaben, 7 freie Schreibaufgaben sowie 5 Schreibminiaufgaben[44] eingesetzt. Bei den Leseaufgaben handelte es sich um Kurztexte von maximal zweiseitiger Länge, darunter sowohl literarische als auch Sachtexte, zu welchen jeweils 5 bis 15 Teilaufgaben gestellt wurden. Bei den Teilaufgaben handelte es sich um (i) Multiple-Choice-Aufgaben, bei denen unter mehreren vorgegebenen Antwortmöglichkeiten die richtige zu wählen war oder in deren Rahmen für mehrere Aussagen Richtigkeit bzw. Falschheit zu bestimmen war, (ii) Zu- und Umordnungsaufgaben, bei welchen beispielweise Textabschnitte vorgegebenen Absatzüberschriften zuzuordnen waren sowie (iii) Fragen, die in Form von Freitext mit einem

[44] Dabei handelt es sich bspw. um Korrektur- oder Überarbeitungsaufgaben, welche für die vorliegende Untersuchung jedoch nicht relevant sind, vgl. hierzu auch Kapitel 3.1.

maximalen Umfang von drei Sätzen zu beantworten waren. Insgesamt wurden 189 Teilaufgaben im Bereich *Lesen* eingesetzt, 151 flossen in die späteren Berechnungen mit ein.[45]

Insgesamt bearbeiteten 1310 Schülerinnen und Schüler jeweils mindestens eine freie Schreibaufgabe sowie drei bis vier Leseaufgaben. Die Leistungswerte dieser Schülerinnen und Schüler bilden die Datenbasis für die im Folgenden vorgestellten Analysen.

7.6.2. Gewinnung der Leistungsrohdaten

Für den Bereich *Schreiben* wurden die Texte aller sieben freien Schreibaufgaben von geschulten Kodiererinnen und Kodierern beurteilt. Diese Beurteilung erfolgte in einem wie in Kapitel 3.3. vorgestellten holistischen System unter Anwendung mehrstufiger Skalen für die Dimensionen *Inhalt*, *Stil* und *sprachliche Richtigkeit* sowie einer Globalskala.

Für den Bereich *Lesen* wurden die Daten dichotom (richtige vs. falsche Antwort) kodiert. Ankreuzaufgaben wurden hierbei über eine automatische Scanerkennungssoftware erfasst. Offene Items wurden von geschulten Kodiererinnen und Kodierern ausgewertet.

7.6.3. Bestimmung der leseschwierigkeitsbestimmenden Merkmale

Für die Stimulus- und Instruktionstexte von sechs der sieben in der Studie eingesetzten Schreibaufgaben wurden zur Bestimmung der sprachlichen Anforderungen an die Lesekompetenz die Ausprägungen der unter 7.4.1. bis 7.4.4. dargelegten Merkmale bestimmt. Eine Aufgabe musste ausgeschlossen werden, da diese mit einem bildlichen Stimulus operierte und nicht sprachlich klassifiziert werden konnte.[46]

[45] 38 Leseitems mussten aufgrund unzureichender Trennschärfe, schlechter Modellpassung, ungenügender Kodierreliabilität oder auch erst durch die Schülerantworten sich offenbarende Unzulänglichkeiten der Aufgabenstellung (z. B. Ambiguitäten) von der Analyse ausgeschlossen werden.

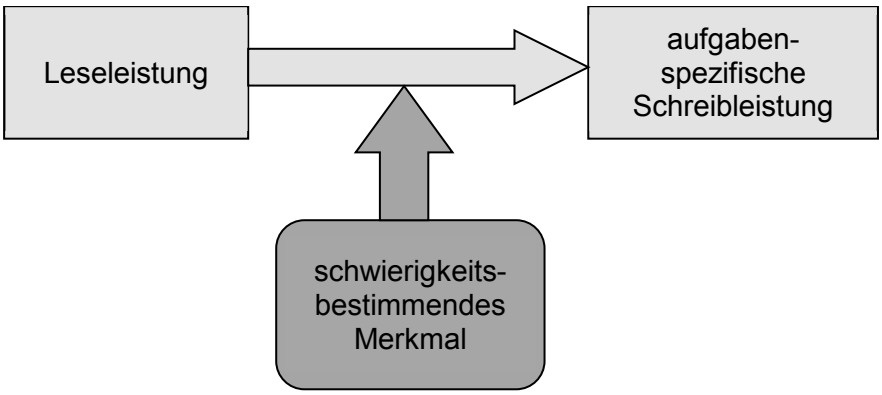
[46] Da nur eine Aufgabe mit einem bildlichen Stimulus operierte, konnte ein Vergleich zwischen textuellen und bildlichen Aufgaben trotz der Relevanz für die Fragestellung nicht vorgenommen werden. Die Bildhaftigkeit ließe sich hier nicht von anderen spezifischen Aufgabenmerkmalen trennen, mögliche Effekte nicht als Effekte der Präsentationsart interpretieren.

7.6.4. Analysen

Zunächst wurden für die Schülerinnen und Schüler die Lese- und Schreibleistungswerte ermittelt, indem die Bewertungen ihrer Antworten in einem zweidimensionalen IRT-Modell mit *ConQuest* (Version 2.0) skaliert wurden. Auf Basis der zweidimensionalen Skalierung wurden Personenparametern in Form von PVs[47] und WLEs[48] ermittelt.

Um zu testen, ob der Zusammenhang zwischen *Lesen* und *Schreiben* in Abhängigkeit von der Leseschwierigkeit des Stimulus der Schreibaufgabe abhängt, wurde für jedes der zwölf in 7.4. vorgestellten Kriterien eine Moderatoranalyse durchgeführt. Dabei diente die Leseleistung in Form des Lesen-WLEs als unabhängige Variable, die aufgabenspezifische Schreibleistung in Form des aufgabenspezifischen Skalenwerts als abhängige Variable, das schwierigkeitsbestimmende Kriterium als Moderatorvariable (vgl. Abbildung 7.6.4.1). Zur besseren Interpretation wurden die Lese- und Schreibleistungswerte *z*-transformiert in die Analyse miteinbezogen.

Abbildung 7.6.4.1: ***Illustration der Moderatoranalyse.***



Als Analyseeinheit dienten die vorliegenden bewerteten Schülertexte. Insgesamt lagen Daten von 2266 Texten von insgesamt 1310 Schülerinnen und Schülern vor.

Die Moderatoranalysen wurden als Mehrebenenanalysen mit der Leseleistung auf Ebene 1, der Textebene,[49] und dem schwierigkeitsbestimmenden Merkmal als steigungsmoderierendem Faktor auf Ebene 2, der Aufgabenebene, durchgeführt. Eine

[47] PV: *plausible value* (Mislevy et al., 1992); vgl. auch Kapitel 3.6.
[48] WLE: *Weighted Likelihood Estimate* (Warm, 1989); vgl. auch Kapitel 3.6.
[49] Da einige Schülerinnen und Schüler zwei, andere nur eine Schreibaufgabe bearbeitet haben, die Zusammenhänge zwischen Lese- und Schreibkompetenz jedoch schreibaufgabenspezifisch erfolgte, entspricht Ebene 1 der Textebene und nicht der Schülerebene. Lese-WLEs lagen auf Schülerebene vor. Schülerinnen und Schüler, welche zwei Schreibaufgaben bearbeitet haben, kommen auf Ebene 1, der Textebene, zweifach im Datensatz mit entsprechendem Lesen-WLE vor.

Mehrebenenanalyse wurde gewählt, da diese Modellierung die Datenstruktur adäquat repräsentiert: Die Varianz in den Merkmalen besteht auf Aufgaben-, nicht auf Textebene; die Texte sind in Aufgaben genestet.

Das jeweilige Mehrebenenmodell wurde als Random-Coefficient-Modell spezifiziert, d. h. sowohl für die Regressionskonstante als auch für die Regressionssteigung wurde eine Zufallskomponente auf Ebene 2 (Aufgabenebene) in das Modell mitaufgenommen:

Ebene 1: $\text{Schreibleistung}_{\text{Aufgabe}} = \beta_0 + \beta_1 \cdot \text{Leseleistung} + r$

Ebene 2: $\beta_0 = \gamma_{00} + u_0$

$\beta_1 = \gamma_{10} + \boldsymbol{\gamma_{11}} \cdot \text{Merkmal} + u_1$

wobei: β_0 = Regressionskonstante (Intercept)

β_1 = Regressionssteigung (Slope)

r = Residuum

Merkmal = Ausprägung des jeweiligen leseschwierigkeitsbestimmenden Merkmals

γ_{00} = Erwartungswert der Regressionskonstanten, wenn *Merkmal* gleich null ist.

u_0 = gruppen-, hier aufgabenspezifische Zufallskomponente der Regressionskonstanten

γ_{10} = Erwartungswert der Regressionssteigung, wenn *Merkmal* gleich null ist.

γ_{11} = Einfluss von *Merkmal* auf β_{1j}

u_1 = gruppen-, hier aufgabenspezifische Zufallskomponente der Regressionssteigung

Der hervorgehobene Koeffizient $\boldsymbol{\gamma_{11}}$ quantifiziert den im Fokus der Fragestellung stehenden Moderatoreffekt.

Die Analyse wurde mit *HLM* (Version 6.0) durchgeführt. Die Merkmale auf Ebene 2 wurden gesamtmittelwertszentriert (*grand centered*) in der Spezifikation des Modells in die Analyse einbezogen. Diese Zentrierung wurde vorgenommen, da für die meisten der Merkmale 0 kein sinnvoller, d. h. real möglicher Skalenwert ist (Krause & Urban, 2013; W. Langer, 2010; Nezlek, Schröder-Abé & Schütz, 2006). Für die Leseleitung musste aufgrund der Einbeziehung *z*-transformierter Daten keine Zentrierung am Gesamtmittelwert vorgenommen werden.

Für jedes Merkmal wurde bestimmt, in welchem Umfang unter Einbeziehung des Merkmals die Residualvarianz u_1 abnimmt im Vergleich zu einem Modell, in welchem dieser Faktor nicht einbezogen wird, dem sogenannten Nullmodell:

Ebene 1: $\text{Schreibleistung}_{\text{Aufgabe}} = \beta_0 + \beta_1 \cdot \text{Leseleistung} + r$

Ebene 2: $\beta_{0} = \gamma_{00} + u_0$

$\beta_1 = \gamma_{10} + u_1$

Der aufgeklärte Varianzanteil berechnet sich als $\Delta var(u_1)/var(u_{1(\text{Nullmodell})})$, wobei $\Delta var(u_1) = var(u_{1(\text{Nullmodell})}) - var(u_{1(\text{Merkmalsmodell})})$.[50]

Bei mehreren statistisch bedeutsamen Merkmalen wurden im Anschluss zum Vergleich dieser Effekte Analysen mit entsprechend mehreren Moderatoren auf Ebene 2 durchgeführt. Dieser Vergleich der Effekte diente zum einen zur Überprüfung, ob es sich um unabhängige Effekte handelt und beispielsweise nicht ein Effekt durch einen anderen getragen wird, zum anderen zur Ermittlung, welchen Anteil das jeweilige Merkmal am Gesamteffekt hat und in welchem Verhältnis dieser Anteil zu dem Anteil anderer Merkmale steht. Dem entsprechenden Modell liegt folgende mathematische Spezifikation zugrunde:

Beispiel für zwei Merkmale:

Ebene 1: $\text{Schreibleistung}_{\text{Aufgabe}} = \beta_0 + \beta_1 \cdot \text{Leseleistung} + r$

Ebene 2: $\beta_{0} = \gamma_{00} + u_0$

$\beta_1 = \gamma_{10} + \boldsymbol{\gamma_{11}} \cdot \text{Merkmal}_A + \boldsymbol{\gamma_{12}} \cdot \text{Merkmal}_B + u_1$

Beispiel für drei Merkmale:

Ebene 1: $\text{Schreibleistung}_{\text{Aufgabe}} = \beta_0 + \beta_1 \cdot \text{Leseleistung} + r$

Ebene 2: $\beta_{0} = \gamma_{00} + u_0$

$\beta_1 = \gamma_{10} + \boldsymbol{\gamma_{11}} \cdot \text{Merkmal}_A + \boldsymbol{\gamma_{12}} \cdot \text{Merkmal}_B + \boldsymbol{\gamma_{13}} \cdot \text{Merkmal}_C + u_1$

Des Weiteren wurden mögliche partielle Effekte (Bortz, 2005; J. Cohen & Cohen, 1975; Urban & Mayerl, 2008) geprüft, indem für alle in der Einzelmerkmalsanalyse nicht bedeutsamen Merkmale jedes dieser Merkmale separat zusammen mit einem oder mehreren der in Einzelmerkmalsanalyse signifikanten Merkmale als Moderatoren in ein Modell mitaufgenommen wurden.[51]

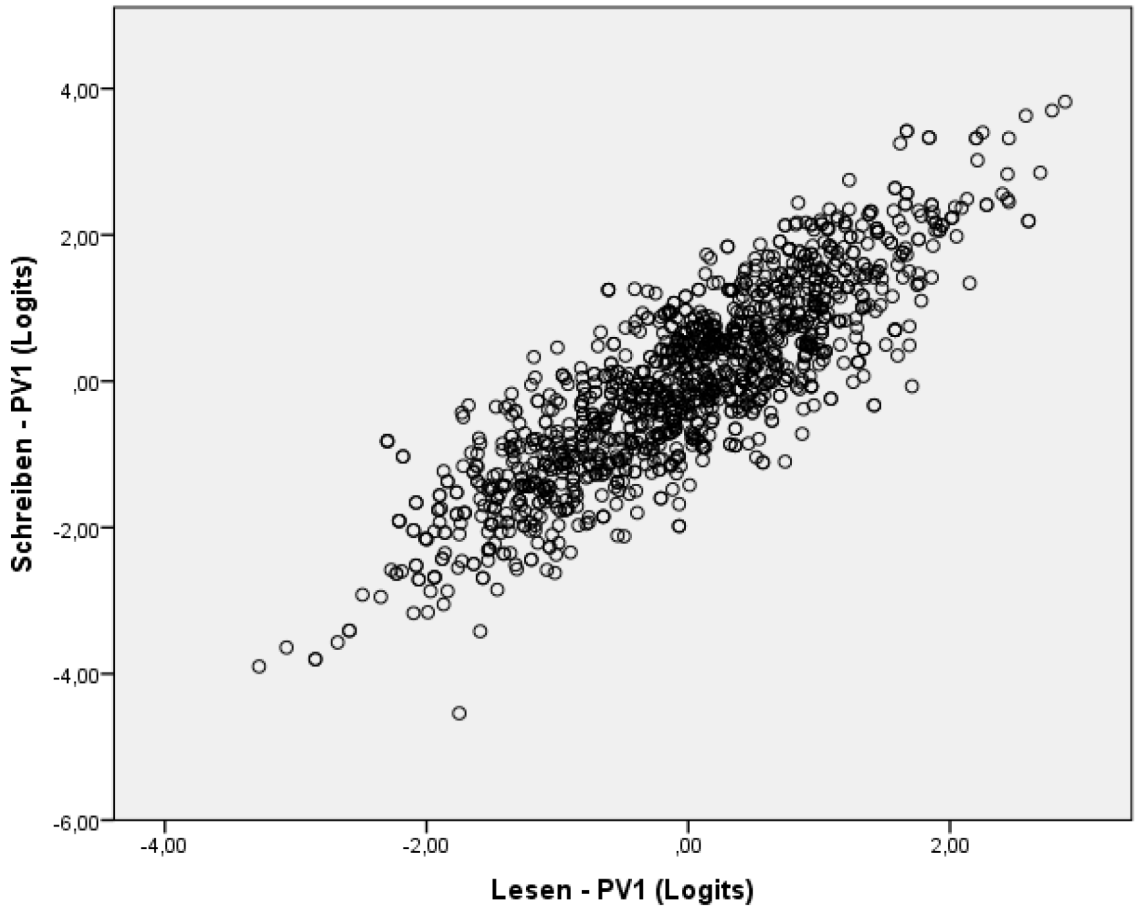
[50] *var* steht für: Varianz.

[51] Dieses schrittweise Vorgehen ist darin begründet, dass eine Einbeziehung mehrerer oder gar aller Faktoren in ein Mehrebenenmodell aufgrund der geringen Anzahl an Ebene-2-Einheiten nicht sinnvoll bzw. ab einer

7.7. Ergebnisse

Die zweidimensionale IRT-Modellierung für *Lesen* und *Schreiben* schätzt eine latente Korrelation von .86 zwischen den beiden Kompetenzen. Die EAP/-PV-Reliabilitäten betragen .74 für *Schreiben* und .83 für *Lesen*. Abbildung 7.7.1 veranschaulicht den Zusammenhang zwischen den beiden Kompetenzen.

Abbildung 7.7.1: ***Streudiagramm: Zusammenhang zwischen Lese- und Schreibkompetenz anhand von Plausiblen Values (PVs); Skaleneinheiten = Logits.***



Prädiktorenzahl von 5 gar nicht möglich war. Darüber hinaus erweisen sich einige der Merkmale als nicht sinnvoll miteinander kombinierbar, da sie sich als Alternativen verstehen (bspw. *Anteil seltener Wörter* und *Anzahl seltener Wörter*) oder andere Maße integrieren (bspw. *LIX* und *Wörter pro Satz*).

Das Nullmodell, welches keines der möglichen moderierenden Merkmale miteinbezieht und den Zusammenhang zwischen der Lesekompetenz der Schülerinnen und Schüler und der aufgabenspezifischen Schreibleistung modelliert, führt zu den in Tabelle 7.7.1 dargestellten Ergebnissen.

Tabelle 7.7.1: ***Nullmodell: Modellierung der Abhängigkeit von aufgabenspezifischen Schreibleistungen mit aufgabenspezifischer Varianz auf Ebene 2.***

	γ_{00}	γ_{10}	var(r)	var(u_0)	var(u_1)
Nullmodell	-0.005	0.508***	0.726	0.017***	0.013***

*** $p < .001$

Das Modell spezifiziert γ_{00}, d. h. die Regressionskonstante ohne aufgabenspezifische Varianz, als nicht bedeutsam von 0 verschieden, was unter Einbeziehung *z*-transformierter Werte und gegebener hoher Korrelation zu erwarten war. Diese hohe Korrelation zeigt sich in der um aufgabenspezifische Varianz befreiten Regressionssteigung γ_{10} von 0.51, welche bedeutsam von Null verschieden ist. Erhöht sich die Leseleistung um eine Standardabweichung, steigt die aufgabenspezifische Schreibleistung um circa eine halbe Standardabweichung. Die aufgabenspezifische Varianz der Regressionskonstante, u_0, welche sich als Indikator für unterschiedliche Aufgabenschwierigkeiten interpretieren lässt, erweist sich als bedeutsam von 0 verschieden, auch wenn nur 2.2 % der Gesamtleistungsvarianz auf Unterschiede im Aufgabenniveau entfallen. Auch auf die aufgabenspezifische Varianz der Regressionssteigung u_1 entfallen insgesamt nur 1.6 % der Gesamtvarianz, was sich allerdings dennoch als bedeutsam von 0 verschieden erweist. Aufgrund dessen, dass aufgabenspezifische Varianz in der Regressionssteigung vorliegt, konnten Folgeanalysen durchgeführt werden und geprüft werden, ob unter Einbeziehung der leseschwierigkeitsbestimmenden Merkmale Teile dieser Varianz erklärt werden können.

Die deskriptiven Kennwerte der leseschwierigkeitsbestimmenden Merkmale sind in Tabelle 7.7.2 dargestellt. Unter Einbeziehung jeweils eines der schwierigkeitsbestimmenden Merkmale als Moderator zeigen sich die in Tabelle 7.7.3 dargestellten Effekte.[52]

[52] Zu Tabelle 7.7.3 sei angemerkt, dass Werte negativer Varianzaufklärung keine Seltenheit bei der Berechnung von Mehrebenenanalysen sind. Sie beruhen auf einem Overfit des Modells (Hox, 2002) und können in ihrer Negativität nicht interpretiert werden, sondern sind wie ein aufgeklärter Varianzanteil von 0 % aufzufassen.

Tabelle 7.7.2: ***Schwierigkeitsbestimmende Merkmale: Mittelwerte, Standardabweichungen und Extrema.***

Merkmal	**Mittelwert**	**Standard-abweichung**	**Minimum°**	**Maximum°**
Textlänge / Textmenge				
Zeichenzahl	1459.17	620.19	708	2479
Silbenzahl	436.00	179.58	219	731
Wortzahl	259.50	114.29	124	457
Satzzahl	28.00	18.13	17	64
sprachliche Komplexität				
Zeichen pro Wort	5.74	0.34	5.48	6.42
Silben pro Wort	1.70	0.11	1.59	1.89
Wörter pro Satz	10.31	3.94	6.28	15.52
lexikalisches Niveau				
mittlere Häufigkeitsklasse	8.02	0.37	7.73	8.64
Anzahl seltene Wörter	14.33	5.54	8	21
Anteil seltene Wörter (%)	8.44	1.97	5.97	10.77
Kombinationsmaße				
LIX	45.17	7.44	35	55
Flesch-Index	52.83	10.91	55	64

° Natürliche Einheiten wurden ohne Nachkommastellen dargestellt.

Tabelle 7.7.3

Ergebnisse der merkmalsspezifischen Zwei-Ebenen-Moderatoranalysen.

Merkmal	γ_{00}	γ_{10}	γ_{11}	**var(r)**	**var(u_0)**	**var(u_1)**	**Δvar(u_1)/ var($u_{1(Nullmodell)}$)**
Textlänge / Textmenge							
Zeichenzahl	-0.005	0.508***	0.000	0.726	0.017***	0.0149***	-0.19
Silbenzahl	-0.005	0.508***	0.000	0.726	0.017***	0.0125***	0.00
Wortzahl	-0.005	0.508***	0.000	0.726	0.017***	0.0152***	-0.22
Satzzahl	-0.005	0.508***	-0.003	0.726	0.017***	0.0119***	0.05
sprachliche Komplexität							
Zeichen pro Wort	-0.005	0.507***	0.176	0.726	0.017***	0.0169***	-0.35
Silben pro Wort	-0.005	0.507***	0.331	0.726	0.017***	0.0170***	-0.36
Wörter pro Satz	-0.005	0.505***	0.024*	0.726	0.017***	0.0011	0.91
lexikalisches Niveau							
mittlere Häufigkeitsklasse	-0.005	0.507***	0.266*	0.725	0.017***	0.0013	0.90
Anzahl seltene Wörter	-0.005	0.507***	0.011	0.726	0.017***	0.0077**	0.38
Anteil seltene Wörter	-0.005	0.508***	0.020	0.726	0.017***	0.0122***	0.02
Kombinationsmaße							
LIX	-0.005	0.508***	0.005	0.726	0.017***	0.0132***	-0.06
Flesch-Index	-0.005	0.507***	0.007	0.726	0.017***	0.0140***	-0.12

* $p < .05$ ** $p < .01$ *** $p < .001$

Erwartungsgemäß werden die Regressionskonstante, die aufgabenspezifische Varianz der Regressionskonstanten sowie die Residualvarianz auf Ebene 1 stabil, d. h. in allen Modellen identisch, geschätzt. Auch die um aufgabenspezifische Varianz befreite Regressionssteigung erweist sich als weitgehend stabil und schwankt lediglich in der dritten Nachkommastelle. Die hinsichtlich einer Moderation getesteten Merkmale erweisen sich weitgehend als nicht bedeutsam, mit Ausnahme der beiden Faktoren *Wörter pro Satz* (im Folgenden auch: *syntaktische Komplexität*) und *mittlere Häufigkeitsklasse* (im Folgenden auch: *Seltenheit der Wörter*). Wenn jeweils nur einer dieser beiden Faktoren der einzig einflussreiche wäre, betrüge die aufgeklärte aufgabenspezifische Regressionssteigungsvarianz 90 bzw. 91 %. Eine Steigerung der syntaktischen Komplexität um 1 Wort pro Satz erhöhte die Regressionssteigung zwischen der Leseleistung und der aufgabenspezifischen Schreibleistung um 0.024. Eine Steigerung der Seltenheit der Wörter um eine Häufigkeitsklasse erhöhte die Regressionssteigung um 0.266. Aufgrund der verschiedenen Skalen der beiden Variablen und der unterschiedlichen Varianzen sind diese Werte nur schwer vergleichend zu interpretieren. Rechnet man die Effekte in Standardabweichungen um, so zeigen sich in etwa vergleichbare Effekte für die beiden Faktoren: Bei Steigerung der syntaktischen Komplexität um eine Standardabweichung erhöht sich die Regressionssteigung um 0.095 (0.024 × 3.94); bei Steigerung der Seltenheit der Wörter um eine Standardabweichung erhöht sich die Regressionssteigung um 0.098 (0.266 × 0.37).

Eine aufgeklärte Varianz von rund 90 % für jeden der beiden Faktoren indiziert bereits, dass die Faktoren wesentliche gemeinsame Varianzanteile besitzen, weshalb beide Faktoren in ein Modell einbezogen wurden. Die Ergebnisse sind in Tabelle 7.7.4 dargestellt.

Tabelle 7.7.4: ***Ergebnisse der Zwei-Ebenen-Moderatoranalyse unter Einbeziehung der Faktoren „mittlere Häufigkeitsklasse“ und „Wörter pro Satz“.***

γ_{00}	γ_{10}	γ_{11}	γ_{12}	var(r)	$var(u_0)$	$var(u_1)$	$\Delta var(u_1)/ var(u_{1(Nullmodell)})$
-0.005	0.507***	0.1919	0.0076	0.725	0.017***	0.001	0.92

γ_{11}: mittlere Häufigkeitsklasse γ_{12}: Wörter pro Satz * $p < .05$ ** $p < .01$ *** $p < .001$

Weder die syntaktische Komplexität noch die Seltenheit der Wörter erweisen sich in diesem Modell als statistisch bedeutsame Einflussfaktoren, obwohl sie gemeinsam 92 % der Steigungsvarianz erklären.[53] Dies deutet darauf hin, dass keiner der beiden Merkmale alleine der tragende Faktor des Effektes ist, sondern beide Faktoren gemeinsam wirksam sind.

Die Analysen auf mögliche partielle Effekte, d. h. alle Analysen, die neben einem oder beiden der in Einzelanalyse signifikanten Merkmale eines der übrigen Merkmale einbeziehen, zeigen keine bedeutsamen Effekte für die in der Einzelanalyse nicht bedeutsamen Merkmale. Es liegen folglich auch keine partiellen Effekte für diese Merkmale vor. Die Faktoren *Wörter pro Satz* und *mittlere Häufigkeitsklasse* zeigen in diesen Analysen stabile und im Vergleich zu den vorgehenden Analysen identische Effekte.[54]

7.8. Zusammenfassung, Diskussion und Ausblick

Die Ergebnisse der Studie zeigen, dass ein hoher Zusammenhang zwischen der Lese- und der Schreibkompetenz von Schülerinnen und Schülern in der achten Jahrgangsstufe besteht. Dieser Zusammenhang ist erwartungskonform, da beim Lesen und Schreiben beteiligte Prozesse zum Teil schriftsprachspezifisch, zum Teil allgemeinsprachspezifisch und somit domänenunabhängig sind.

Die Ermittlung des Zusammenhangs zwischen *Lesen* und *Schreiben* diente als Grundlage für die Ausgangsfragestellung der vorgestellten Teilstudie, inwiefern bei der Messung von Schreibkompetenzen (schreibkompetenzirrelevante) Lesefähigkeiten miterfasst werden. Aufgrund der gemeinsamen Kompetenzanteile, welche auf identischen kognitiven Sprachverarbeitungsprozessen beruhen, ist ein gewisser Lesekompetenzanteil gerade nicht konstruktirrelevant. Aspekte, die jedoch einerseits einen Teil der Lesekompetenz bestimmen, andererseits nicht als Teil des vorliegenden Schreibkompetenzkonstrukts angesehen werden können, sind sprachliche Aspekte, die die Sprachrezeptionsschwierigkeit bedingen. Im Rahmen der vorliegenden Arbeit wurde der Fokus auf diese sprachlichen Aspekte gelegt und

[53] Das Vorliegen einer extrem hohen Varianzaufklärungsrate bei gleichzeitiger Nichtsignifikanz lediglich zweier Einflussfaktoren beruht darauf, dass lediglich 6 Aufgaben (Ebene-2-Einheiten) in die Analyse eingingen. Signifikanztests sind in starkem Maße fallzahlabhängig, bei geringer Fallzahl werden nur extrem starke Effekte als signifikant ausgewiesen. Andererseits ist bei niedrigerer Fallzahl eine höhere Varianzaufklärung zu erwarten, da mit abnehmenden Fällen sich das geschätzte Modell immer stärker einem deterministischen Modell annähert (vgl. auch nachfolgende Anmerkungen unter 7.8.).

[54] Insgesamt handelt es sich hierbei um 30 mögliche, davon 22 inhaltlich sinnvolle Modelle. Aufgrund der hohen Anzahl der Modelle bei einheitlichem Befundmuster wurde auf eine ausführliche Darstellung dieser Ergebnisse hier verzichtet. Die Ergebnisse befinden sich jedoch im Anhang unter Tabelle A 7.7.1 bis A.7.7.3.

konkret die Fragestellung verfolgt, ob der Zusammenhang zwischen *Lesen* und *Schreiben* in Abhängigkeit der sprachlichen Rezeptionsschwierigkeit der Schreibaufgabeninstruktion (und des entsprechenden Stimulus) variiert. Es zeigten sich hierbei generell nur sehr schwache Effekte. Nur 3.8 % der gesamten Schreibleistungsvarianz entfiel auf Aufgabenebene, davon 58 % auf Ebene der Aufgabenschwierigkeit (= 2.2 % der Gesamtleistungsvarianz) und 42 % auf Ebene der Zusammenhangsstärke (= 1.6 % der Gesamtleistungsvarianz).

Die Zusammenhangsstärke wurde hierbei in Abhängigkeit mehrerer Faktoren berechnet, welche die Leseschwierigkeit des Vortextes der jeweiligen Schreibaufgabe quantifizierten und einen oder mehrere Aspekte der Bereiche *lexikalisches Niveau*, *sprachliche Komplexität* und *Textmenge* erfassten. Es erwiesen sich insgesamt zwei dieser Faktoren als bedeutsam, die durchschnittliche Häufigkeit/Seltenheit der im Text enthaltenen Wörter und die syntaktische Komplexität des Textes.

Sprachliche Komplexität, darunter spezifisch die syntaktische Komplexität, und lexikalisches Niveau sind auch diejenigen Faktoren, welche in der Leseschwierigkeitsforschung als die beiden zentralen Größen dargestellt werden (Best, 2006a, DuBay, 2004; Ernst, 2011b; Meuers, 2014; Mihm, 1973). In der vorliegenden Studie zeigte die syntaktische Komplexität als einziges sprachliches Komplexitätsmaß bedeutsame Effekte. Sprachliche Komplexitäten, welche auf Wortebene anzusiedeln sind, wiesen keinerlei Effekte auf. Dies ist erwartungskonform, da sich diese auf basale Prozesse der Worterkennung beziehen, welche in diesem Stadium der Sprachentwicklung in der Regel keine Hürde mehr darstellen (Fitzgerald & Shanahan, 2000; Frith, Wimmer & Landerl, 1998; Stanovich, 1980).

Hinsichtlich des lexikalischen Niveaus besteht in der Leseschwierigkeitsforschung Konsens über dessen Relevanz, jedoch existiert kein einheitliches Maß, um dieses zu messen (vgl. die Ausführungen unter 7.4.3.). In der vorliegenden Studie zeigte sich die durchschnittliche Häufigkeitsklasse der Wörter als einflussreicher Faktor, während weder die Anzahl noch der Anteil seltener Wörter in geeigneter Weise differenzierten, um lexikalische Schwierigkeitseffekte zu quantifizieren. Da im Rahmen dieser Studie jedoch lediglich ein moderierender Einfluss untersucht wurde und keine direkten Effekte des Faktors auf Aspekte der Leseleistung, müssen weitere Studien, welche die Leseschwierigkeit von Texten unmittelbar untersuchen, prüfen, ob dieser Faktor generell ein gut geeigneter ist, um das lexikalische Niveau zu quantifizieren.

Auch die beiden populären leseschwierigkeitsbestimmenden Maße LIX und der Flesch-Index integrieren jeweils die Aspekte *syntaktische Komplexität* und *lexikalisches Niveau* in ihrer

jeweiligen Formel, allerdings konnte weder für den LIX, welcher zumindest mittelbar den Anteil seltener Wörter einbezieht, noch für den Flesch-Index, welcher ebenfalls nur mittelbar die Häufigkeit/Seltenheit aller Wörter einbezieht, ein Effekt gefunden werden. Neben anderen Kritikpunkten, wie etwa, dass die konkreten Formeln, vor allem die festen Faktoren in diesen, angepasst werden müssten (Mikk, 1995; Mikk & Elts, 1999), besteht auch insofern Kritik an diesen Maßen, dass sie lexikalisches Niveau nur indirekt über Wortlängen erfassen (vgl. Kapitel 7.4.4.): Auch wenn Wortlänge und Worthäufigkeit miteinander korrelieren, so bestimmt die Wortlänge die Worthäufigkeit nicht direkt und somit auch nicht hinreichend adäquat (Ernst, 2011a; Klare, 1963; Mikk, 2000). Der Zusammenhang zwischen den beiden Größen scheint in vielen Kontexten nicht ausreichend, um Effekte der Häufigkeit anhand der Wortlänge zu erfassen. Auch in der hier vorgestellten Studie wurden Wortlängenmaße geprüft; für diese konnten jedoch keine bedeutsamen Effekte gefunden werden. Das Ausmaß, in welchem Worthäufigkeit und Wortlänge nicht assoziiert sind, könnte auch eine mögliche Ursache dafür sein, dass Maße wie der LIX oder der Flesch-Index häufig nicht mit subjektiven Schwierigkeitsbeurteilungen von Texten übereinstimmen (Kercher, 2010). Da inzwischen umfangreiche Wörterbücher und Datenbanken vorliegen, in denen die Häufigkeitsklassen von Wörtern einfach und komfortabel abrufbar sind, wie das auch für diese Studie verwendete Wortschatzportal der Universität Leipzig oder die Kookkurrenz-datenbank des Instituts für deutsche Sprache in Mannheim (Perkuhn, Keibel & Kupietz, 2012) ist es wünschenswert, neue Maße zur Erfassung von *Textschwierigkeit* zu generieren und zu erproben, um *lexikalisches Niveau* geeignet zu quantifzieren.

Auch erwies sich in der vorliegenden Studie kein Textlängenmerkmal als bedeutsam. Dies mag darauf zurückzuführen sein, dass die jeweilige Aufgabenbearbeitungszeit inklusive der Rezeptionszeit des Stimulus nur 20 Minuten betrug. Aufgrund dieses Settings lag nur eine begrenze Textlängenvarianz in den Aufgaben vor (17–64 Sätze, 127–457 Wörter, 708–2479 Zeichen). Nold und Rossa (2007) weisen darauf hin, dass Textlängeneffekte im Rahmen von Kurztexten, als welche die Aufgabenstimuli hinsichtlich ihres Umfangs zu kategorisieren sind, (noch) nicht zum Tragen kommen.

Zurückkommend auf die Frage, inwiefern die schwierigkeitsbestimmenden Merkmale die aufgabenspezifischen Unterschiede des Zusammenhangs zwischen Lese- und Schreib-kompetenz erklären, so zeigte sich, dass die beiden bedeutsamen Moderatoren (*mittlere Häufigkeitsklasse der Wörter* und *Wörter pro Satz*) zusammen 92 % der aufgabenspezifischen

Zusammenhangsvarianz aufklärten; die Residualvarianz sank unter Einbeziehung dieser beiden Faktoren in den statistisch nichtbedeutsamen Bereich.

Für die vorliegende Studie lässt sich festhalten, dass ein statistisch bedeutsamer Einfluss der Lesekompetenz bzw. des hier untersuchten Anteils der Lesekompetenz auf die gemessene Schreibkompetenz (über den Basiszusammenhang hinaus) feststellbar ist. Dieser Einfluss ist jedoch sehr gering und klärt lediglich 1.5 % der Gesamtvarianz auf, sodass man von einer minimalen Erfassung konstruktirrelevanter Varianz ausgehen kann, die praktisch jedoch nicht relevant ist. Die Validität des Tests zur Erhebung von Schreibkompetenzen bzw. der entsprechenden Testergebnisse ist/sind nicht in praktisch bedeutsamem Maße durch aufgabenspezifisch unterschiedlich starke Miterfassung von sprachlichen Lesekompetenz-anteilen gefährdet. Dies trifft zumindest auf ein Aufgabenspektrum zu, wie es in der vorliegenden Studie eingesetzt wurde. Diese Aufgaben wurden jedoch auch speziell von erfahrenen Lehrkräften und Fachdidaktikern zum Einsatz als Testaufgabe in einer Schreibkompetenzerhebung am Ende der Sekundarstufe I entwickelt, daher orientieren sich die Aufgaben an einem bestimmten Sprachentwicklungsstand sowie an gegebenen einheitlichen Kontextfaktoren (Bearbeitungszeit, Testsituation etc.) und weisen daher nur eine eingeschränkte Varianz bezüglich der textschwierigkeitsbestimmenden Parameter auf. Inwiefern diese eingeschränkte Varianz praktisch relevant ist, ist jedoch fraglich, da davon auszugehen ist, dass auch im Rahmen anderer Schreibkompetenztests die Aufgaben auf die Zielpopulation und die Testsituation ausgerichtet und abgestimmt von erfahrenen und kundigen Personen entwickelt werden.

Abschließend sei angemerkt, dass eine Generalisierbarkeit der hier vorgestellten Ergebnisse nur eingeschränkt gegeben ist. In die Mehrebenenanalysen flossen lediglich sechs Aufgaben, d. h. sechs Ebene-2-Einheiten ein. Zumindest für inferentielle Aussagen wird häufig eine Mindestanzahl an Ebene-2-Einheiten von 10–30, für Cross-Level-Effekte (wie die Moderatoreneffekte im Rahmen der vorliegenden Studie) oftmals sogar eine höhere Anzahl (50–100) gefordert (Hox, 1998; Kreft, 1996; Maas & Hox, 2004; Snijders & Bosker, 1994). Allerdings weisen einige Autoren darauf hin, dass die festen Effekte, d. h. die Parameter für die Regressionskonstanten (*Intercepts*) und -steigungen (*Slopes*), weitgehend auch bei kleinerer Gruppenzahl hinreichend gut geschätzt werden, die Residualvarianzen hingegen oftmals verschätzt, zumeist unterschätzt, werden (Bell, Morgan, Schoeneberger, Loudermilk, Kromrey & Ferron, 2010; Nezlek et al., 2006; van der Leeden & Busing, 1994). Eine Übertragbarkeit der Befunde auf andere Kontexte sollte daher nicht ungeprüft angenommen

werden, vor allem nicht, wenn sich die Kontexte von dem vorliegenden deutlich, beispielsweise in der Testweise oder dem Aufgabenmaterial unterscheiden. Zukünftige Studien mit einer erhöhten Anzahl an textuellen Schreibaufgaben und einer höheren Varianz in den Ausprägungen der schwierigkeitsbestimmenden Merkmale werden zeigen müssen, ob und in welchem Maße sich die hier gefundenen Effekte verallgemeinern lassen.

8. Halo-Effekte bei der inhaltlichen und stilistischen Textbeurteilung aufgrund der sprachlichen Richtigkeit (Teilstudie III)

In diesem Kapitel wird der Fragestellung nachgegangen, ob und inwiefern eine Beurteilung der inhaltlichen und stilistischen Textqualität anhand der in der Normierungsstudie eingesetzten semiholistischen Skalen unabhängig von der sprachlichen Richtigkeit der Texte gelingt oder ob und inwiefern inhaltliche und stilistische Urteile durch die Sprachrichtigkeit verzerrt sind. Darüber hinaus wird untersucht, ob bei Vorliegen einer Verzerrung, diese einseitig, d. h. nur unter Vorliegen einer gesteigerten Fehlerhaftigkeit der Texte zutage tritt oder ob auch Fehlerfreiheit die Urteile positiv beeinflusst. Des Weiteren wird geprüft, inwiefern vorliegende Verzerrungen von Fehlerzahl, Fehlertypen und Textmerkmalen wie Länge und Komplexität sowie dem zugrunde liegenden Textmuster abhängen.

8.1. Urteilsverzerrungen, Halo-Effekte und ihre Relevanz

Bei Urteilen, die sich auf komplexe Gegenstände beziehen, wie etwa im Rahmen der Textbeurteilung konnten eine Reihe von verzerrenden Einflüssen nachgewiesen werden. Für den schulischen Kontext zeigten diverse Studien, dass die Beurteilung von Schüleraufsätzen mittels Schulnoten nicht (immer) objektiv erfolgt, sondern von diversen Oberflächenmerkmalen des Textes wie orthografischen und grammatikalischen Fehlern (Birkel, 2003; Birkel & Birkel, 2002) oder der Qualität der Handschrift (Briggs, 1970; Sprouse & Webb, 1994) beeinflusst ist. Ferner konnten personenspezifische Verzerrungseffekte beispielsweise aufgrund der Beliebtheit der Schülerinnen und Schüler oder deren Aussehen festgestellt werden (Behrens & Krelle, 2011; Ingenkamp, 1971; Leppert, 2010; Valtin, 2002).

Solche Verzerrungseffekte nennt man Halo-Effekte. Der Begriff *Halo-Effekt* geht auf Thorndike (1920) zurück und bezeichnet das Phänomen, dass Personen bei ihrer Einschätzung anderer Personen bezüglich bestimmter Eigenschaften auf bekannte, dominante oder offensichtliche Eigenschaften zurückgreifen, auch wenn das zu beurteilende Merkmal und das zur Beurteilung herangezogene Merkmal nicht in unmittelbarer Beziehung zueinander stehen. Diese Wahrnehmungsverzerrungen finden unwillkürlich und unbewusst statt. Halo-Effekte treten nicht nur bei der Beurteilung von Personen und Personeneigenschaften auf, sondern können auch auf Produkte dieser Personen – wie etwa geschriebene Texte – übertragen werden.

Im Rahmen eines Schreibassessments wie der hier zugrunde liegenden Normierungsstudie sind Urteilsverzerrungen ausgehend von Personenmerkmalen aufgrund der Anonymität der Auswertung irrelevant, Halo-Effekte auf Basis von Textmerkmalen wie etwa der sprachlichen Richtigkeit sind jedoch theoretisch nicht ausgeschlossen, auch wenn im Rahmen von Schreibassessments andere Bedingungen vorliegen als im schulischen Kontext. Der wichtigste Unterschied zwischen der Aufsatzbeurteilung im Unterricht und der Kompetenzeinschätzung anhand von Texten im Rahmen von Assessmentstudien ist neben der Anonymität der Beurteilung und somit auch deren Unabhängigkeit von personenbezogenen Leistungserwartungen, dass die Beurteilenden, um die für das Assessment notwendigen hinreichend reliablen Urteile zu erzielen, stärker (in Form von detaillierten Manualen) instruiert und im Umgang mit diesen Manualen mehrfach geschult werden.

Wie bereits in zurückliegenden Kapiteln dieser Arbeit dargestellt, erweist sich die Dimension der sprachlichen Richtigkeit theoretisch wie empirisch als unabhängig von den Dimensionen *Inhalt* und *Stil* (Böhme et al., 2009; A. Neumann, 2007; vgl. Kapitel 2.2.1., 3.3., 6.6., 6.7.). Für eine möglichst valide Messung – frei von konstruktirrelevanter Varianz – inhaltlicher und stilistischer Schreibfähigkeiten ist somit eine durch Aspekte der sprachlichen Richtigkeit unverzerrte Messung dieser Dimensionen zu gewährleisten.

8.2. Bisherige Studien und Befunde

8.2.1. Befunde im Rahmen von Qualitäts- und Leistungsbeurteilungen

Studien bezüglich der Fragestellung, wie die sprachliche Richtigkeit bzw. Fehlerhaftigkeit von Texten die Beurteilung dieser Texte auf einer anderen Ebene beeinflusst, liegen für das Deutsche für Texte von Schülerinnen und Schülern der Sekundarstufe I im Rahmen von Schreibassessments bisher keine vor. Bisherige Studien, die der Fragestellung nachgingen, sind weitgehend aus dem englischen Sprachraum und beziehen sich auf den schulischen oder universitären Unterrichtskontext. In einigen dieser Studien wurde eine Fehlertypendifferenzierung vorgenommen. Insgesamt zeigt sich jedoch ein uneinheitliches Befundmuster.

Keine Effekte fand Chase (1968), der die inhaltliche Textbeurteilung von fehlerfreien und fehlerhaften Texten kontrastierte. Letztere wiesen eine relativ hohe Fehlerrate von einem Fehler auf acht Wörter auf. Eine Fehlertypendifferenzierung wurde im Rahmen der Studie nicht vorgenommen, es handelte sich ausschließlich um Lautbuchstabenzuordnungsfehler.

Auch Russell und Tao (2004) konnten keinen Effekt im Kontrast fehlerfreier und fehlerhafter Texte bei der Beurteilung von *Themenentwicklung*, worunter Aspekte wie Ideenentfaltung, inhaltliches Arrangement, inhaltliche Gliederung und Detailreichtum fallen, nachweisen. Bei den Fehlern handelte es sich um alle originären Fehler der Textautoren und somit um Fehler aller Art in unkontrollierten Anteilen.

Birkel und Birkel (2002; auch Birkel, 2003) hingegen fanden in einem regressionsanalytischen Analyseansatz zur Auswertung von Benotungen von Schülertexten mit wenigen vs. vielen Fehlern, wobei die orthografischen Leistungen explizit nicht in die Benotung mit einfließen sollte, dass sich 7 % der Varianz in der Notengebung auf die Fehlerhaftigkeit zurückführen lassen, was die Autoren als beachtlichen, nicht zu vernachlässigenden Effekt interpretieren. Eine Fehlertypisierung wurde im Rahmen dieser Studie nicht vorgenommen.

Fehlertypabhängige Befunde zeigten die Studien von Marshall (1967) und Marshall und Powers (1969), in welchen bei der inhaltlichen Textbeurteilung zwischen Zeichensetzungs-, Lautbuchstabenzuordnungs- und grammatikalischen Fehlern unterschieden wurde; die Fehlerzahl wurde zwischen 0, 6, 12 und 18 (Marshall, 1967) bzw. 0 und 18 (Marshall & Powers, 1969) variiert. Bedeutsam schlechtere Textbeurteilungen zeigten sich bei grammatikalischen und Lautbuchstabenzuordnungsfehlern bei höherer Fehleranzahl (12, 18 bzw. 18).

Auch Linn, Klein und Hart (1972) fanden fehlertypabhängige Effekte bei der Unterscheidung von a) Lautbuchstabenzuordnungsfehlern, b) grammatikalischen Fehlern, c) Zeichensetzungsfehlern, d) Konstruktionsfehlern und Inkonsistenzen im Sprachgebrauch. Es zeigten sich signifikante negative Korrelationen zwischen Fehlerzahl und Textbeurteilung für grammatikalische Fehler (b) sowie für Konstruktionsfehler und Inkonsistenzen im Sprachgebrauch (d).

Bei den angeführten Studien handelt es sich vorwiegend um kleinere Studien aus dem Bereich der Unterrichtsforschung, welche erhebliche Unterschiede im Untersuchungsdesign

aufweisen, so etwa im Textmaterial (natürliches vs. künstlich erstelltes Textmaterial[55]), in Art und Anzahl der Textautoren, in der Text- und Aufgabenmenge, in der Kontrolle und Varianz der Fehleranzahl, in der Authentizität bzw. Manipulation der Fehler, in Art (Profession, Qualifikation) und Anzahl der Beurteilenden, im konkreten Beurteilungskriterium sowie in der Verwendung von Instruktionen und Manualen (vgl. Tabelle 8.2.1.1).

Tabelle 8.2.1.1: ***Studien zur Beurteilung der Qualität von Aufsätzen, eingesetzte Texte, Beurteiler und Materialien.***

Studie	Sprache	Textmaterial	Textautoren	Textmenge	Beurteiler (*N*)	Relevantes Beurteilungs-kriterium	Kodiermanual
Birkel & Birkel (2002) / Birkel (2003)	Deutsch	natürliche Aufsätze, die anschließend in zwei Versionen (viel vs. wenige Fehler) transformiert wurden.	Schüler (4. Jahrgangs-stufe, Deutschland)	4 Aufsätze	89 Grundschul-lehrer	Aufsatzqualität	nein (ungelenkte Notengebung)
Chase (1968)	Englisch	künstlich erstellt	(k.A.)	2 Aufsätze (2 Aufgaben à 1 Antwort)	Studenten (*N* = k.A.)	Inhaltliche Beurteilung	50 % der Rater mit, 50 % ohne Manual
Linn, Klein & Hart (1972)	Englisch	natürliche Aufsätze (anschließend transkribiert)	Jura-Examenskan didaten (USA)	19 verschiedene thematische Einheiten (insgesamt 607) aus 79 Examens-arbeiten	4 Jura-Professoren	Gesamtbenotung	allgemeine Benotungs-hinweise
Marshall (1967) / Marshall & Powers (1969)	Englisch	ein natürlicher inhaltlich sehr guter Text als Antwort auf eine Frage in Amerikanische Geschichte, der inhaltlich künstlich um eine Notenstufe verschlechtert wurde	Schüler (12. Jahrgangs-stufe, USA), Lehrer (*Amerika-nische Geschichte*) zur inhaltlichen Textmani-pulation	1 Text	700 Lehrer (*Amerikanische Geschichte*) / 420 Lehramts-studenten (*Amerikanische Geschichte* und „*Educational Psychology*“)	rein inhaltliche Beurteilung	ja
Russel & Tao (2004)	Englisch	natürlich (in Fehlervariante mit allen originären Fehlern), im Anschluss transkribiert	Schüler (8. Jahrgangs-stufe, USA)	60 Texte	12 Rater (11 davon „middle and high school teachers“)	*Themen-entwicklung*	ja

Im Rahmen dieser Tabelle wurde zugunsten der Übersichtlichkeit auf das sprachliche Gendern verzichtet. Personenbezeichnungen stehen im generischen Maskulin und schließen weibliche Personen ein.

[55] Die Unterscheidung *natürlich* vs. *künstlich* verweist hier auf den Unterschied, ob es sich um authentisches Textmaterial von Schülerinnen und Schülern bzw. Studentinnen und Studenten handelt oder ob die Texte von einem Autor gezielt für die Untersuchung verfasst worden sind.

Darüber hinaus stellt die Tatsache, dass die Studien in unterschiedlichen Sprachen durchgeführt wurden, eine mögliche Quelle für differente Befundmuster dar. Die meisten Untersuchungen wurden im englischen Sprachraum mit englischem Sprachmaterial durchgeführt. Inwieweit die Ergebnisse sich auf das Deutsche übertragen lassen, ist fraglich. Gerade bei Fehlerkategorisierungen wie *spelling errors* oder *phonological errors* sind die sprachlichen Voraussetzungen der Sprachen Englisch und Deutsch sehr verschieden. So finden sich im Deutschen deutlich mehr regelhafte Lautbuchstabenzuordnungen als im Englischen (Caravolas, 2004; Caravolas & Landerl, 2010; Landerl & Wimmer, 2000; Treutlein, 2011; H. Wimmer, 1993; Ziegler, Perry & Coltheart, 2000). Auch lassen die verschiedenen Grammatiken der beiden Sprachen eine Übertragbarkeit von Effekten von grammatikalischen Fehlern fraglich erscheinen. Beispielsweise ist das Deutsche morphologisch komplexer, das Englische syntaktisch rigider (W. Abraham, 1995; Roelcke, 2003). Damit liegt den Sprachen ein anderer Pool an Fehlermöglichkeiten zugrunde, d. h. die potentiellen Gelegenheiten, Fehler zu machen, sind sprachabhängig andere. Dies wiederum kann dazu führen, dass bei der Textbeurteilung andere Fehlertoleranzen entstehen oder andere Fehlergewichtungen von bestimmten grammatikalischen Fehlern vorgenommen werden.

Ein weiterer zentraler Aspekt, in welchem sich die bisherigen Studien von der Textbeurteilung im Rahmen von Large-Scale-Schreibassessments unterscheiden, ist der Umstand, dass im Rahmen von umfangreichen Assessments die Beurteilenden mehrfach in der Anwendung der Kodiermanuale geschult werden. Diese Schulungen dienen im Rahmen von Large-Scale-Kompetenzmessungen der Gewährleistung der Reliabilität der Urteile und somit gerade der Minimierung eventueller subjektiver Verzerrungsfaktoren.

Eine Untersuchung, die im Rahmen eines Large-Scale-Schreibassessments durchgeführt wurde, welche hinsichtlich der zugrunde liegenden Sprache, der Natürlichkeit und Authentizität der Texte und der Fehler, der verwendeten Beurteilungsinstrumente sowie der Beurteilergruppe der hier vorgestellten Untersuchung gleicht, ist die bereits in Kapitel 7 herangezogene Normierungsstudie des IQB im Kompetenzbereich *Schreiben* für die Primarstufe. Böhme et al. (2009) untersuchten anhand der Daten dieser Studie Halo-Effekte bei der Beurteilung von Texten von Grundschülerinnen und Grundschülern. Diese Texte wurden von geschulten Kodiererinnen und Kodierern anhand der holistischen und semiholistischen Skalen (*Global*, *Inhalt*, *Stil*, *sprachliche Richtigkeit*) in einer für die Primarstufe angepassten Version beurteilt. Die Autoren konnten mittels eines von Bechger, Maris und Hsiao (2007) vorgestellten statistischen Verfahrens Halo-Effekte zwischen den

Beurteilungsebenen nachweisen. Es zeigten sich zwischen allen textspezifischen Urteilen auf den vier Skalen zur Textbeurteilung (*Global*, *Inhalt*, *Stil*, *sprachliche Richtigkeit*) substantielle Halo-Effekte quantifizierende Korrelationen; diese Korrelationen betrugen .20 für *Stil* und *sprachliche Richtigkeit* sowie .23 für *Inhalt* und *sprachliche Richtigkeit* (S. 321). Bei diesen Halo-Effekten quantifizierenden Korrelationen bleibt jedoch offen, welches die Ursachen dieser Halo-Effekte sind. Sie indizieren lediglich, ob Texte, die beispielsweise raterspezifisch stilistisch besser (als von anderen Ratern) bewertet wurden, auch von diesem Rater orthografisch-grammatisch besser bewertet wurden, geben jedoch keinerlei Auskunft über kausale Zusammenhänge. Auch ist die Übertragbarkeit der Ergebnisse auf die Beurteilung von Texten von Schülerinnen und Schülern am Ende der Sekundarstufe I fraglich. So erweisen sich zum einen die Strukturen von sprachlichen Kompetenzen über Entwicklungsstadien hinweg nicht notwendigerweise als stabil (vgl. Kapitel 6.7., 7.2., 7.8; Jude, 2008), insofern raterseitig implizites oder explizites Wissen über entwicklungsspezifische Fähigkeitszusammenhänge vorliegt, könnte dies möglicherweise Halo-Effekte entsprechend stärken oder schwächen. Zum anderen muss davon ausgegangen werden, dass die Wahrnehmung von Verletzungen der sprachlichen Richtigkeit abhängig von der Kenntnis des konkreten Entwicklungsstadiums der Schreibenden ist; so werden Rechtschreibfehler von Grundschülerinnen und Grundschülern möglicherweise nachsichtiger wahrgenommen als solche von Erwachsenen.

8.2.2. Befunde im Rahmen von Personen- und Produkteinschätzungen

Neben Studien zur Schreibleistungs- und Textqualitätsbeurteilung finden sich solche aus dem Bereich der Sozialpsychologie, in deren Rahmen die Probanden Eigenschaften und Fähigkeiten anderer Personen anhand von Schreibprodukten oder Eigenschaften dieser Schreibprodukte selbst einschätzen sollten. Die sozialpsychologische Einbettung dieser Untersuchungen lassen die Ergebnisse in einem tieferen explanatorischen Zusammenhang interpretieren als dies im Rahmen obiger Studien möglich war.

Kreiner, Schnakenberg, Green, Costello und McClin (2002) ließen Psychologiestudenten (N = 82) den Autor eines Film-Reviews, von welchem gemäß Fehleranzahl (0, 4, 12) und Fehlerart (phonologische Fehler; Tippfehler) verschiedene Varianten erstellt wurden, hinsichtlich der Ausprägung der Fähigkeiten *Schreibkompetenz*, *Intelligenz* und *logisches Denkvermögen* beurteilen: Es zeigte sich, dass *Schreibkompetenz* bei höherer Fehleranzahl niedriger

eingestuft wurde als unter Fehlerabsenz. Phonologische Fehler führten tendenziell zu einer stärkeren Abstufung gegenüber der fehlerfreien Textvariante als Tippfehler. Ein analoges, aber schwächeres Muster zeigte sich für *Intelligenz*. Die Einschätzung von *logisches Denkvermögen* blieb unbeeinflusst. Die Autoren interpretierten dies als Indiz dafür, dass die zugrunde liegende Fähigkeit (*Schreibkompetenz)* und weitere hierarchisch distalere Metafähigkeiten (wie etwa *Intelligenz)* (in Abhängigkeit von der hierarchischen Nähe/Distanz) beeinflusst sind, unbetroffene Fähigkeiten (wie etwa *logisches Denkvermögen*) werden nicht beeinflusst.

Figueredo und Varnhagen (2005) legten zwei von einem Studenten verfasste Aufsätze 173 Psychologiestudenten zur Beurteilung unter anderem der folgenden Fähigkeiten des Autors vor: *Rechtschreibfähigkeit*, *Schreibfähigkeit*, *allgemeine Intelligenz* und *mathematische Fähigkeit*. Die Texte wurden in drei Varianten vorgelegt: ohne Fehler, mit Homophon-Fehlern[56] und mit Nichthomophon-/Pseudohomophon-Fehlern[57]; die Fehler wurden an 15 verschiedenen Wörtern (sowie allen folgenden Wiederholungen dieser Wörter) realisiert. *Rechtschreibfähigkeit*, *Schreibfähigkeit* und *allgemeine Intelligenz* wurden gemäß dieser Reihenfolge mit abnehmender Stärke unter der „keine Fehler"-Bedingung bedeutsam besser beurteilt als unter den Fehlerbedingungen; Pseudohomophon-Fehler führten tendenziell zu einer stärkeren Abstufung als Homophon-Fehler. Die Beurteilung der *mathematischen Fähigkeit* blieb von Fehlerpräsenz und Fehlerart unbeeinflusst.

Varnhagen (2000) ließ Schülerinnen und Schüler unterschiedlicher Jahrgangsstufen diverse Aussagen über Texte und Textautoren in Abhängigkeit der Fehlerhaftigkeit der Texte beurteilen. Es zeigten sich Effekte der sprachlichen Richtigkeit beispielsweise für *Rechtschreibkompetenz des Auto*rs, die *zu erwartende Schulnote des Textes*, die *Eignung als Mathematiknachhilfelehrer* oder die *Sorgfältigkeit des Autors*, nicht jedoch für die *Interessantheit des Textes,* den *Unterhaltungswert des Textes* oder die *mathematischen Fähigkeiten des Autors*.

Sowohl Varnhagen (2000) als auch Figueredo und Varnhagen (2005) interpretieren diese Ergebnisse als Unterscheidungsfähigkeit der Urteilenden, relatierte und unrelatierte Fähigkeiten, d. h. Fähigkeiten die in einem (objektiven) Zusammenhang zueinander stehen und solche, die in keinem solchen Zusammenhang stehen, zu unterscheiden.

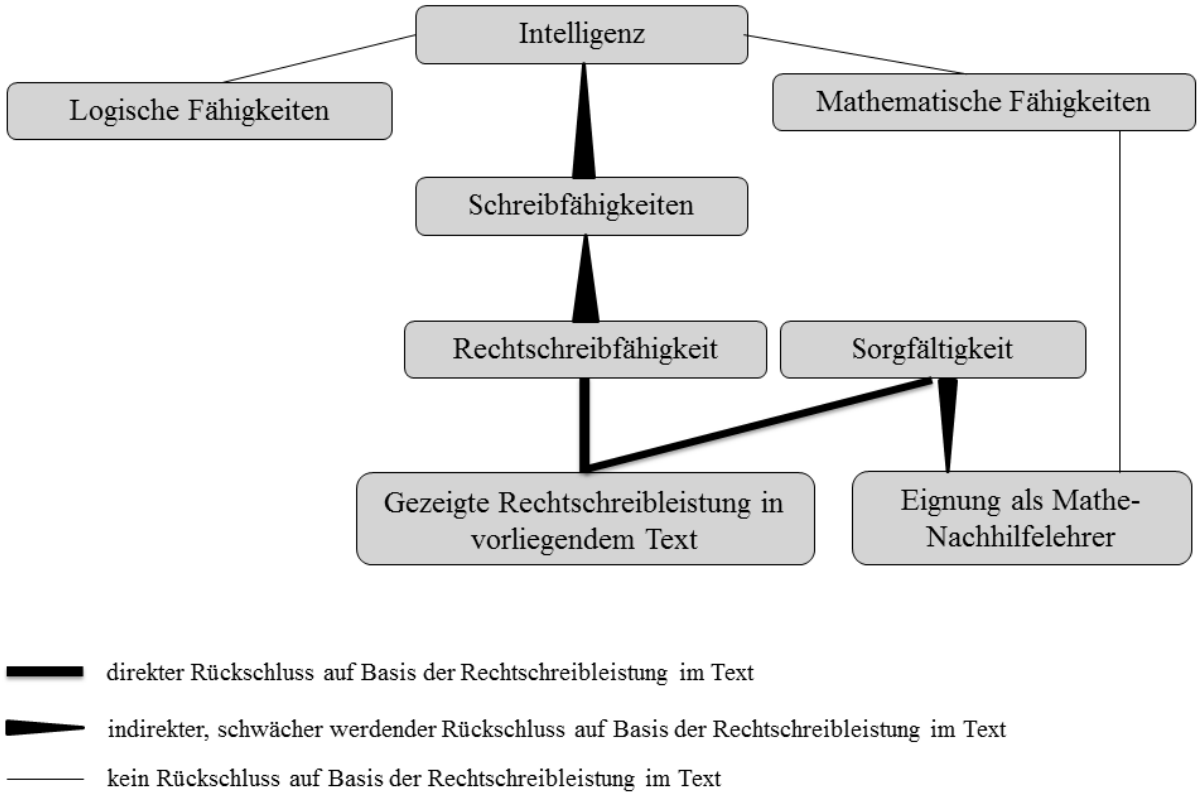
---

[56] Homophone sind gleichlautende, aber anders geschriebene Wörter einer Sprache, bspw. *Ferse* und *Verse* im Deutschen.

[57] Pseudohomophone sind in einer Sprache nicht existierende Wörter, die es allerdings gemäß Buchtstaben- und Lautabfolgeregeln der Sprache geben könnte, so bspw. *Fogel*.

Die Schlussfolgerungen und Interpretationen der Autoren der drei hier angeführten Studien implizieren somit Annahmen über das mentale Modell der Beurteilenden, welche Personenfähigkeiten wie miteinander in Bezug stehen, welche Fähigkeiten sich als Teilfähigkeiten anderer Fähigkeiten erweisen und auch in welchem Abstand (in der Hierarchietiefe) die Fähigkeiten zueinander stehen. Diese Interpretationen der Autoren ermöglichen es, ein Modell aufzustellen, welches die Annahmen der Beurteilenden über Beziehungen zwischen Fähigkeiten der Autoren darstellt (vgl. Abbildung 8.2.2.1).

Abbildung 8.2.2.1: ***Exemplarische Modellierung der Annahmen der Urteilenden über die Zusammenhänge der Fähigkeiten und Eigenschaften von Beurteilten ausgehend von der sprachlichen Richtigkeit eines vorliegenden Textes.***



Urteile erstrecken sich nach diesem Modell auf unmittelbar relatierte Fähigkeiten (z. B. *Rechtschreibfähigkeit* oder *Sorgfältigkeit*). Mittelbare Fähigkeiten, etwa Unterfähigkeiten einer gemeinsamen zugrunde liegenden Fähigkeit oder daraus abgeleitete Eigenschaften (z. B. *Eignung als Mathematik-Nachhilfelehrer*) sowie hierarchiehöhere Fähigkeiten (z. B. *Intelligenz*) werden in Abhängigkeit von ihrer wahrgenommenen Nähe/Distanz mitbeurteilt. Dieses Modell liefert somit eine mögliche explanatorische Basis, warum ggf. bestimmte Halo-Effekte auftreten, andere möglicherweise nicht.

8.2.3. Spezifische sprachliche Aspekte

8.2.3.1. stilistische Textbeurteilung

Untersuchungen, welche die ausschließliche Beurteilung des Stils in Abhängigkeit von der sprachlichen Richtigkeit der entsprechenden Texte betrachten, liegen bislang keine vor. Die in 8.2.1. und 8.2.2. dargelegten Studien fokussierten entweder rein inhaltliche Textqualitäts-aspekte oder die Textqualität als Ganzes (unter Ausschluss orthografisch-grammatischer Aspekte). Der Stil kann jedoch als Bindeglied zwischen der Sprache und dem Inhalt angesehen werden – so wird Stil in Conrad (1995) definiert als „durch die Auswahl aller sprachlichen Mittel charakterisierte mündliche oder schriftliche Verwendungsweise der Sprache. Die Auswahl ist vom Zweck abhängig, und die Kombination unterliegt den Regeln der betreffenden Sprache, der Stilistik selbst und historischen gesellschaftlichen Veränderungen.“ (S.231), somit besteht eine engere konzeptionelle Nähe zwischen Stil und Sprache als zwischen Inhalt und Sprache.

8.2.3.2. Textbeurteilung in Abhängigkeit von Textmustern

Einige Untersuchungen konnten zeigen, dass die Textsorte bzw. das Textmuster eines Textes (*argumentieren*, *beschreiben*, *erzählen* etc.) und die damit verbundene Art der Aufbereitung der Information Einflüsse auf kognitive Kapazitäten wie das Arbeitsgedächtnis oder die Schlussfolgerungsleistung des Lesers hat (Baretta, Tomitch, MacNair, Lim & Waldie, 2009; Berkowitz & Taylor, 1984; Falke, 2008; Taylor & Beach, 1984; Sáenz & Fuchs, 2002; Wolfe & Mienko, 2007). Zur Fragestellung, ob das Textmuster Auswirkungen auf die Textbeurteilung hat, liegen bisher weder theoretische Annahmen noch empirische Befunde vor.

8.2.3.3. Textbeurteilung in Abhängigkeit von Fehlertypen

Die Studien von Marshall (1967) und Linn et al. (1972) lieferten Evidenz für Urteilsverzerrungen bei Vorliegen grammatischer Fehler, nicht jedoch anderer Fehlertypen wie Zeichensetzungsfehler oder Fehler auf graphematischer Ebene. Eine mögliche Erklärung für diesen Effekt liefern Erkenntnisse aus der psycholinguistischen Grundlagenforschung. Liegen grammatische Fehler in einem Text vor, so ist der Leser gezwungen, grammatische

Reparaturen und Reanalysen durchzuführen. Vor allem bei syntaktischen Fehlern sind diese Reparaturen und Reanalysen mit gesteigerten Verarbeitungskosten verbunden (u. a. Frazier, 1979; Frazier & Rayner, 1982). Aufgrund der Bindung kognitiver Kapazitäten für diese grammatischen Korrekturen kann für den Textbeurteilungsprozess von einer Verlagerung zugunsten unbewusster Verarbeitungsstrategien ausgegangen werden.

8.2.3.4. Textbeurteilung in Abhängigkeit von Textlänge und Textkomplexität

Länge und Komplexität gelten als zwei sprachliche Eigenschaften von Texten, welche die Rezeptionsschwierigkeit mitbestimmen (vgl. Kapitel 7.4.). Mit steigender Textlänge und vor allem mit zunehmender Textkomplexität werden ansteigend mehr kognitive Ressourcen seitens des Lesers und somit auch des Beurteilenden in Anspruch genommen (Köster, 2005; Nunan & Koepke, 1995; Schweitzer, 2007). Dies kann in einer dementsprechend vermindert kontrollierten Beurteilungsaktivität resultieren (Baddeley, 1986; Lund, 2001), die eine verstärkte Aktivierung unbewusster Verarbeitungsprozesse mit sich bringt.

8.3. Fragestellungen und Hypothesen

Die Gesamtuntersuchung besteht aus insgesamt fünf Untersuchungsteilen, die sukzessive präsentiert werden.

In Untersuchungsteil I wird den folgenden Fragestellungen nachgegangen:

a) Gibt es Halo-Effekte und somit unerwünschte Verzerrungen der Inhaltsurteile aufgrund der sprachlichen Richtigkeit der entsprechenden Texte? – Bisherige Untersuchungen liefern hierzu ein uneinheitliches Befundmuster (vgl. 8.2.1). Aufgrund dessen, dass es sich bei den Beurteilenden in der vorliegenden Studie um mehrfach in der Unterscheidung der Schreibdimensionen geübte und geschulte Raterinnen und Rater handelt, wird davon ausgegangen, dass es keine Verzerrungen der Inhaltsurteile aufgrund der sprachlichen Richtigkeit gibt.

b) Gibt es Halo-Effekte und somit unerwünschte Verzerrungen der Stilurteile aufgrund der sprachlichen Richtigkeit der entsprechenden Texte? – Bezüglich dieser Fragestellungen liegen bisher keinerlei Untersuchungen vor, die eine Erwartungsrichtung indizieren würden. Da der Stil jedoch, wie unter 8.2.3.1. als Bindeglied zwischen Sprache und Inhalt fungiert, wird

angenommen, dass, insofern Halo-Effekte bei der Inhaltsbeurteilung zutage treten, auch Halo-Effekte bei der Stilbeurteilung auftreten und dass die Halo-Effekte für die Stilbeurteilung größer sind als diejenigen bei der Beurteilung des Inhalts. Sollten, wie vermutet, keine Halo-Effekte für die Inhaltsurteile vorliegen, können keine eindeutigen Vermutungen hinsichtlich des Vorliegens von Halo-Effekten für die Stilurteile angestellt werden. Der Fragestellung wird somit weitgehend explorativ nachgegangen.

c) Zeigen sich textmusterspezifische Unterschiede bezüglich der möglichen Halo-Effekte aus a) und b)? – Wie in 8.2.3.2. dargelegt, liegen zur Beantwortung dieser Frage bisher weder theoretische Annahmen noch empirische Befunde vor. Aufgrund der potentiellen Möglichkeit der Auswirkung der textmusterspezifischen kognitiven Anforderungen, wurden die Analysen zu den Hauptfragestellungen a) und b) textmusterspezifisch durchgeführt; die Fragestellung c) wird explorativ verfolgt.

Untersuchungsteil II widmet sich der Fragestellung, ob, falls eine Verzerrung vorliegt, diese nur unter Präsenz von Fehlerhaftigkeit auftritt oder ob auch eine positive Beeinflussung bei weitgehender Fehlerfreiheit stattfindet.

Bisherige Studien haben sich auf korrelative oder gruppenvergleichende Analysen beschränkt. Im Rahmen dieser Analysen ist es nicht möglich, zwischen einer reinen Beeinflussung unter Fehlerpräsenz und einer Beeinflussung unter Fehlerpräsenz (negativ) sowie Fehlerabsenz (positiv) zu unterscheiden. Durch die Anwendung mess- und strukturmodellierender statistischer Verfahren ist es jedoch möglich, die beiden Annahmen zu kontrastieren und zu vergleichen. Zur Beantwortung dieser Fragestellung wird aufgrund fehlender theoretischer und empirischer Anhaltspunkte ein exploratives Vorgehen gewählt.

Die Untersuchungsteile III und IV beschäftigen sich mit folgender Fragestellung: Falls eine Verzerrung vorliegt, hängt diese von der Anzahl und der Art der Fehler ab?

Da Halo-Effekte per definitionem von Eigenschaften ausgehen, die dominanter und dem Beurteilenden präsenter sind, wird erwartet, dass der Grad der Verzerrung mit der Anzahl der Fehler steigt. Dies stünde im Einklang mit den Befunden von Marshall (1967), Marshall und Powers (1969) und Kreiner et al. (2002). Hinsichtlich des Auftretens von Halo-Effekten in Abhängigkeit bestimmter Fehlertypen legen die Studien von Marshall (1967) und Linn et al. (1972) sowie die Ausführungen zur schriftlichen Sprachrezeption (vgl. 8.2.3.3.) nahe, dass grammatikalische Fehler zu einer (höheren) Verzerrung führen, von Zeichensetzungsfehlern

sollten keine Effekte ausgehen; für Fehler auf graphematischer Ebene lässt sich aufgrund der uneinheitlichen Ergebnisse bisheriger Untersuchungen keine eindeutige Vermutung ableiten.

Untersuchungsteil V widmet sich der Fragestellung, inwieweit der Grad möglicher Verzerrungen von Textkomplexität und Textlänge abhängt. Bisher liegen diesbezüglich keine systematischen Untersuchungen vor. Aufgrund der gesteigerten Bindung kognitiver Ressourcen bei längeren und vor allem komplexeren Texten und der damit verbundenen höheren Wahrscheinlichkeit der Aktivierung unbewusster Verarbeitungsprozesse (vgl. 8.2.3.4.) und somit auch der Entstehung von Halo-Effekten, wird eine gesteigerte Verzerrung möglicherweise bei längeren[58] und vor allem bei komplexeren Texten erwartet.

8.4. Untersuchungsteil I: Mögliche Halo-Effekte bei der Textbeurteilung aufgrund der sprachlichen Richtigkeit

8.4.1. Methoden

8.4.1.1. Aufgaben und Textauswahl

Für die Untersuchung wurde eine Auswahl von sechs der zwölf in der Normierungsstudie eingesetzten (vgl. Kapitel 3.1. & 3.4.) Aufgaben getroffen, je zwei pro Textmuster: zwei narrative, zwei argumentierende und zwei informierende, i. e. die beiden beschreibenden Aufgaben. Für diese sechs Aufgaben wurden die jeweils 60 bis 90 auf der Skala *sprachliche Richtigkeit* am schlechtesten beurteilten Schülertexte herangezogen (10 bis 20 % der Gesamtmenge der orthografisch-grammatisch bewerteten Texte). Die Textgrundlage bestand aus insgesamt 430 Texten von 405 Schülerinnen und Schülern.

8.4.1.2. Erstellung des Experimentalmaterials

Die Schülertexte wurden in zweifacher Weise transkribiert, zum einen in fehlerbeibehaltender Form, zum anderen in fehlerbereinigter Form. Die Fehlerbereinigung erfolgte gemäß einem zuvor entwickelten detaillierten Fehlerkorrekturmanual. Korrigiert wurden:

[58] Inwiefern der Faktor Länge im gegebenen Setting von 1-3seitigen handschriftlichen Texten bereits Effekte evoziert, ist fraglich (vgl. auch die Ergebnisse aus Kapitel 7.7. und entsprechende Interpretationen unter Kapitel 7.8.).

- alle orthografischen Fehler: Fehler auf graphematischer Ebene (inklusive s-ss-ß-Schreibung), Groß-/Kleinschreibung, Getrennt-/Zusammenschreibung, Silbentrennung, Apostrophierung,
- Zeichensetzungsfehler aller Art,
- wort- und einfache satzgrammatische Fehler, d. h. rein grammatische Fehler, insofern keine stilistischen Aspekte durch die Korrektur mitverändert wurden, so etwa Wortbildungsfehler, Verletzungen der grammatischen Kongruenz zwischen Subjekt und Prädikat, Deklinations- und Konjugationsfehler oder Wortstellungsfehler.

Nicht korrigiert aufgrund der möglichen Mitveränderung stilistischer Aspekte wurden:

- lexikalische Fehler (falsche Wortwahl),
- Satzbaufehler aufgrund fehlender Satzglieder,
- falsch gewählte Verknüpfungsmittel,
- die Wahl des falschen Tempus.

8.4.1.3. Textbeurteilung

Die Transkripte wurden von jeweils vier Kodierenden, die ebenfalls im Rahmen der Normierungsstudie als Kodierende tätig waren, auf ihre inhaltliche und stilistische Qualität anhand der entsprechenden Skalen beurteilt. Die Kodiererzuteilung erfolgte gemäß einem zuvor entwickelten ausbalancierten Design, sodass jeder Kodierende alle Schülertexte kodierte, die eine Hälfte in der fehlerkorrigierten, die andere Hälfte in der fehlerbelassenen Variante. Jede Textvariante wurde von zwei Kodierenden beurteilt. Kodiererpaarungen waren über alle möglichen sechs paarweisen Kombinationen ausgeglichen (vgl. Tabelle 8.4.1.3.1).

Die Interrater-Reliabilität für die beiden Skalen *Inhalt* und *Stil* aller sechs Aufgaben bewegte sich im zufriedenstellenden bis guten Bereich (vgl. Tabelle 8.4.1.3.2).

Tabelle 8.4.1.3.1: ***Testdesign zur Beurteilung einer Aufgabe mit 60 Schülertexten.***

	Kodierer 1	Kodierer 2	Kodierer 3	Kodierer 4
Texte 1–10	f	f	k	k
Texte 11–20	f	k	f	k
Texte 21–30	f	k	k	f
Texte 31–40	k	f	f	k
Texte 41–50	k	f	k	f
Texte 51–60	k	k	f	f

f: Beurteilung der fehlerhaften Variante; k: Beurteilung der korrigierten Variante

Tabelle 8.4.1.3.2: ***Intraklassenkorrelationen für Aufgaben und Skalen.***

Aufgabe	**ICC, Inhalt**	**ICC, Stil**
Informierend I	.79	.62
Informierend II	.65	.62
Argumentierend I	.75	.64
Argumentierend II	.64	.55
Narrativ I	.72	.66
Narrativ II	.64	.65

8.4.1.4. Analysen

Zunächst wurde aufgrund des Zusammenhangs zwischen den Schreib(teil)kompetenzen (vgl. Kapitel 6.6.) im Rahmen einer Vorabanalyse geprüft, ob durch die Auswahl der orthografisch-grammatischen am niedrigsten bewerteten Texte das zur Verfügung stehende mögliche Abweichungsspektrum eingeschränkt ist; dies wäre beispielsweise der Fall, wenn diese Auswahl auch zu einer Selektion der inhaltlich und/oder stilistisch am schlechtesten bewerteten Texte geführt hätte, sodass Aufstufungen bei zufälliger Variation im Kodierverhalten wahrscheinlicher sind als Abstufungen. Diese Überprüfung erfolgte im Rahmen einer Wahrscheinlichkeitsbestimmung von möglichen Auf- und Abstufungen sowie Nichtveränderungen für die Inhalts- und Stilurteile, d. h. jedem Urteil (bei der Bewertung der

fehlerhaften Texte) wurden rechnerisch die Wahrscheinlichkeiten für eine positive Veränderung, für eine negative Veränderung und für das Ausbleiben einer Veränderung zugewiesen.[59] Diese Analyse diente lediglich der Kontrolle, ob durch eine eventuell schiefe Basisverteilung der Beurteilungen der fehlerhaften Texte höhere Urteile für die fehlerfreien Texte bereits bei beliebigem Urteilsverhalten erwartbar gewesen wären.

Anschließend wurden für die Hauptanalyse die Mittelwerte aus den Urteilen der beiden Kodierenden für jeden Text in jeder Textvariante (fehlerbelassen und fehlerkorrigiert) berechnet. Daraufhin wurden die Abweichungen aller Textpaare, d. h. die jeweiligen Differenzen zwischen dem Kodiermittelwert der fehlerkorrigierten Variante und dem Kodiermittelwert der fehlerbelassenen Variante bestimmt. Diese Differenzen wurden mittels Wilcoxon-Vorzeichen-Rang-Tests auf Verschiedenheit von Null überprüft (Wilcoxon, 1945); Effektstärken wurden in Form von Cliff's Delta (für abhängige Stichproben) bestimmt (Cliff, 1993; Macbeth, Razumiejczyk & Ledesma, 2011). Diese Analysen wurden textmuster-spezifisch durchgeführt. Zum Vergleich der Textmuster wurden Kruskal-Wallis-Tests berechnet (Kruskal & Wallis, 1952). Als Effektstärke wurde Cliff's Delta (für unabhängige Stichproben) herangezogen.

8.4.2. Ergebnisse und Diskussion

Die abgegebenen Inhalts- und Stilurteile erstrecken sich in jeweils beiden Beurteilungs-varianten (*fehlerhaft* und *korrigiert*) über das komplette Skalenspektrum mit einer höheren Besetzung der mittleren Kategorien gegenüber den Randkategorien (vgl. Abbildungen 8.4.2.1. bis 8.4.2.4).

[59] So wurde etwa ein Urteil der Stufe 1 mit einer Nichtveränderungswahrscheinlichkeit von 0.25 (eines von vier möglichen Folgeurteilen, in diesem Fall: Stufe 1) und einer Aufstufungswahrscheinlichkeit von 0.75 (drei von vier möglichen Folgeurteilen, in diesem Fall: Stufe 2, Stufe 3 und Stufe 4) versehen; ein Urteil der Stufe 2 wurde mit einer Wahrscheinlichkeit von je 0.25 für Abstufung und Nichtveränderung und Aufstufungs-wahrscheinlichkeit von 0.5 versehen. Für Urteile der Stufe 3 und 4 wurden nach analogem Prinzip die entsprechenden Wahrscheinlichkeiten erfasst.

Abbildung 8.4.2.1: ***Relative Häufigkeiten: inhaltliche Beurteilung der fehlerhaften Texte.***

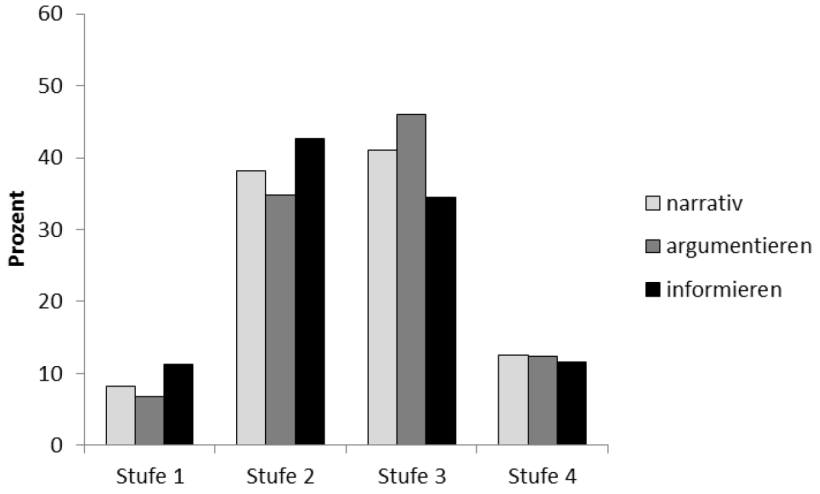


Abbildung 8.4.2.2: ***Relative Häufigkeiten: inhaltliche Beurteilung der korrigierten Texte.***

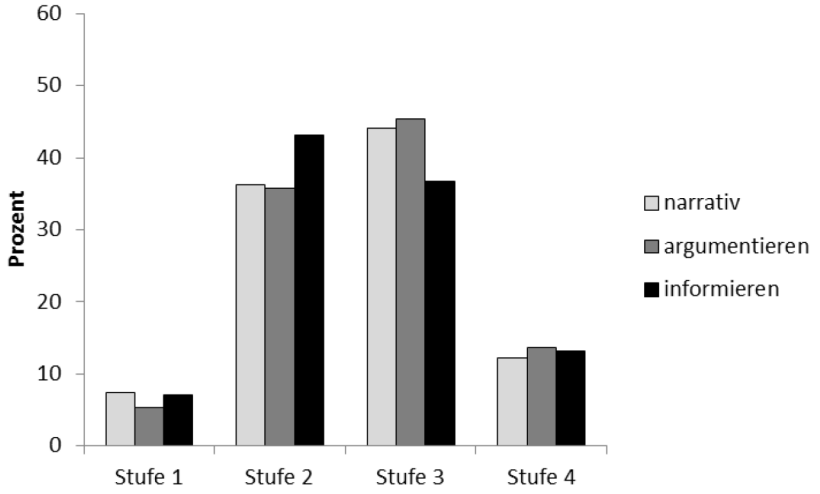


Abbildung 8.4.2.3: ***Relative Häufigkeiten: stilistische Beurteilung der fehlerhaften Texte.***

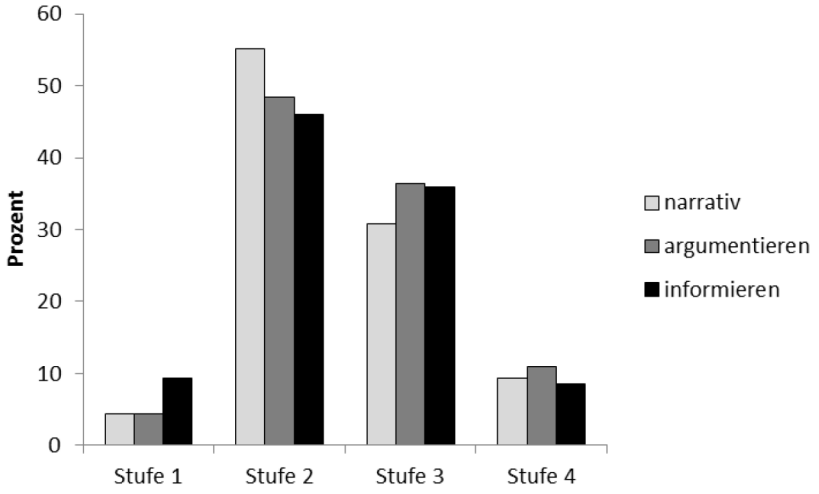
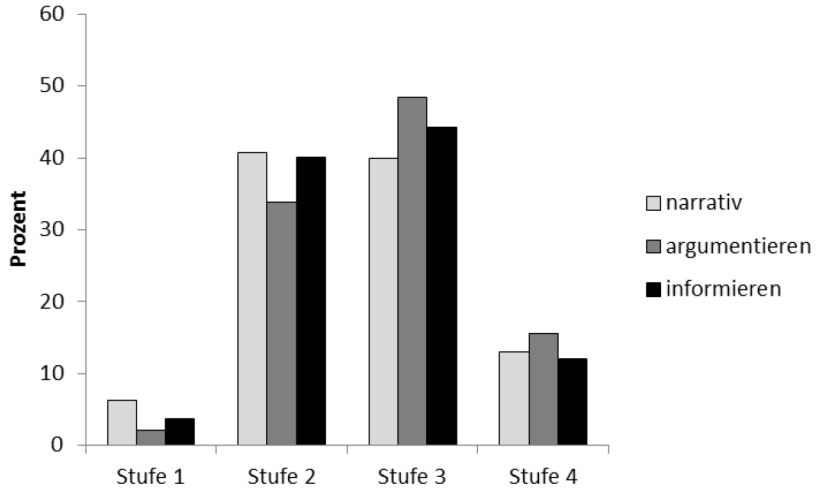


Abbildung 8.4.2.4: ***Relative Häufigkeiten: stilistische Beurteilung der korrigierten Texte.***



Dabei zeigt sich, dass die Verteilungen des Subsamples im Vergleich zum Gesamtsample (vgl. Kapitel 4.3.1.) nach links verschoben sind; vor allem die Stufe 2 ist stärker, die Stufe 3 schwächer besetzt. Dies ist aufgrund der Assoziation zwischen den einzelnen Schreibkompetenzdimensionen erwartungskonform. Die Verteilungen des Subsamples resultieren in weniger schiefen Verteilungen, was sich positiv auf die Folgeuntersuchungen auswirkt, da sich somit die Wahrscheinlichkeiten für mögliche Auf- und Abstufungen (ausgehend von den fehlerhaften Texten) in etwa die Waage halten. Diese Wahrscheinlichkeiten sind in den Abbildungen 8.4.2.5 und 8.4.2.6 dargestellt.

Abbildung 8.4.2.5: ***Wahrscheinlichkeiten für Aufstufungen, Abstufungen und Nichtveränderungen im inhaltlichen Urteil ausgehend von der fehlerhaften Textvariante.***

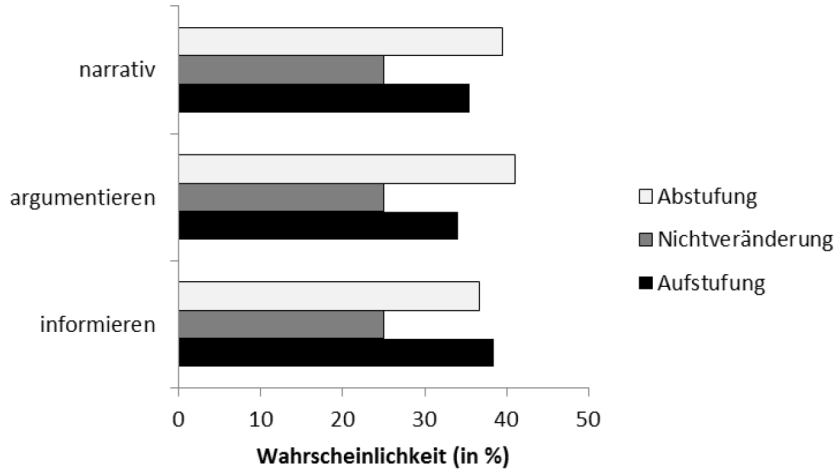
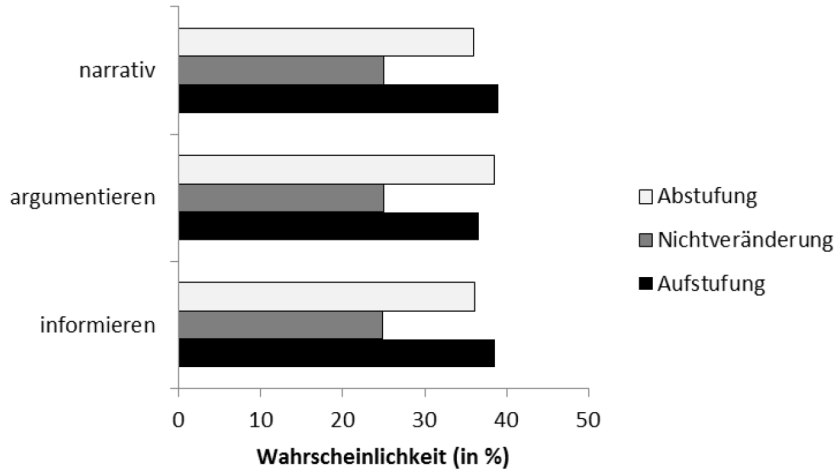


Abbildung 8.4.2.6: ***Wahrscheinlichkeiten für Aufstufungen, Abstufungen und Nichtveränderungen im stilistischen Urteil ausgehend von der fehlerhaften Textvariante.***

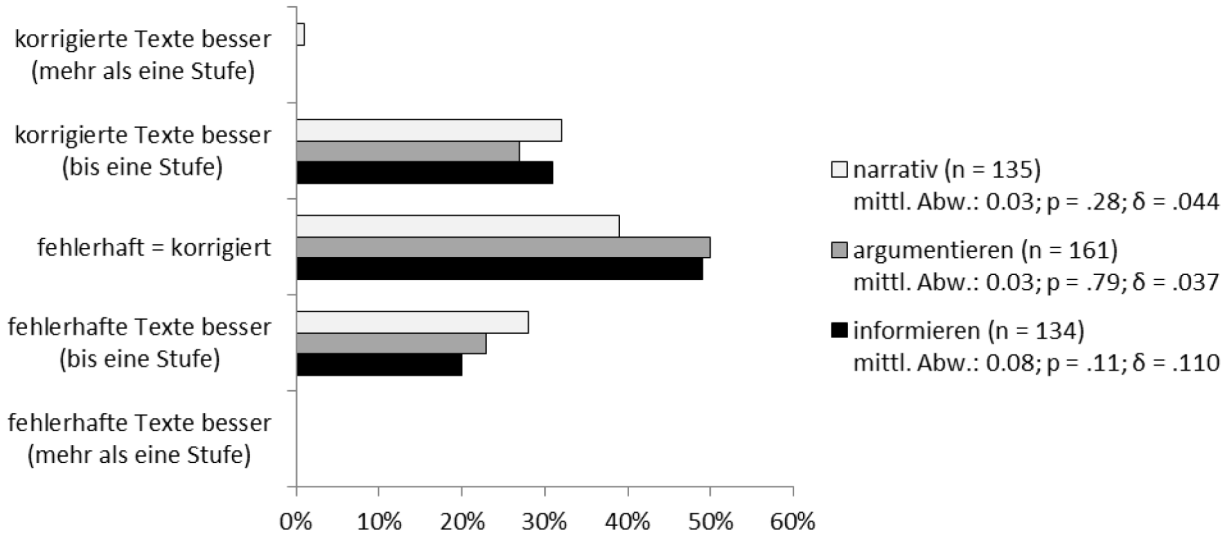


Es zeigt sich für beide Beurteilungsdimensionen und alle drei Textmuster eine annähernd gleich hohe Wahrscheinlichkeit für mögliche Auf- und Abstufungen ($\Delta_{|Auf\text{-}Ab|} < 5$ %), lediglich für das argumentierende Textmuster liegt für die inhaltliche Beurteilung mit 7 % Differenz eine gesteigerte Abstufungswahrscheinlichkeit (41 %) gegenüber der Aufstufungswahrscheinlichkeit (34 %) vor. Wie die weiteren Ergebnisse zeigen werden, erweist sich dieser Unterschied jedoch als irrelevant.

Hinsichtlich einer möglichen Verzerrung der Inhaltsurteile aufgrund der sprachlichen Richtigkeit zeigen sich für alle drei Textsorten keine bedeutsamen Effekte. Die mittleren Abweichungen zwischen korrigierten und fehlerbelassenen Textvarianten betragen für die

narrativen, argumentierenden und informierenden Texte zwischen 0.03 und 0.08 Stufen (alle $p > .10$). Die Effektstärken indizieren, dass keine Effekte vorliegen: Cliff's $\delta = .044$ (*narrativ*), .037 (*argumentieren*), .110 (*informieren*).[60] Abbildung 8.4.2.7 illustriert die relative Verteilung der Texte gemäß ihrer Eigenschaft, ob beide Textvarianten gleich bewertet wurden oder ob eine der beiden Textvarianten besser beurteilt wurde. Abweichungen wurden gemäß ihrer Stärke (leichte vs. starke Abweichung) unterschieden. Abweichungen im Umfang größer als eine Stufe sollten nur unter systematischen Verzerrungsaspekten auftreten. Bei Abweichungen bis zu einer Stufe sind systematische Abweichungen mit erwartbaren Kodiererschwankungen konfundiert (vgl. Kapitel 3.5., Ausführungen und Ergebnisse zur exakten und näheren prozentualen Übereinstimmung). Da man davon ausgehen kann, dass es keine systematische Verzerrung derart gibt, dass fehlerhafte Textvarianten besser bewertet werden als korrigierte, dient die Kategorie *fehlerhafte Texte bis eine Stufe besser* als Vergleichswert für die erwartbaren nicht bedeutsamen Kodiererschwankungen für die Kategorie *korrigierte Texte bis eine Stufe besser*.

Abbildung 8.4.2.7: ***Relative Verteilung der Textpaare gemäß ihrer inhaltlichen Beurteilung.***

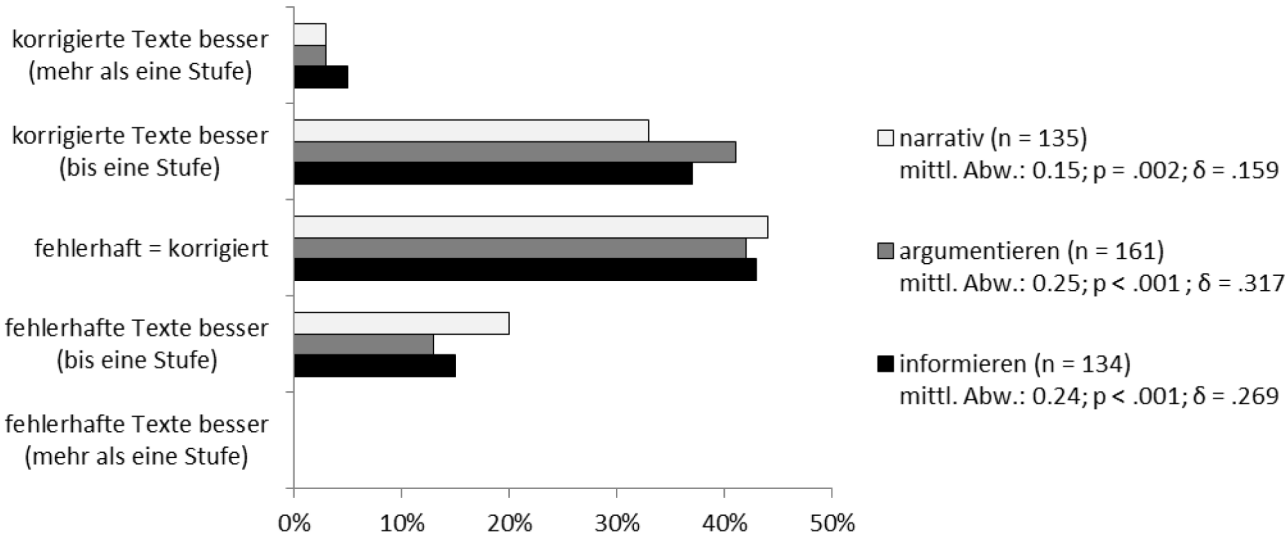


Es zeigen sich keine textmusterspezifischen Unterschiede der Verteilungen ($p = .411$; $\delta_{arg./inf.} = .031$; $\delta_{arg./nar.} < .001$; $\delta_{inf./nar.} = .073$).

[60] Bei Cliff's Delta handelt es sich um eine Effektstärke für ordinale Daten, welche mit dem bekannteren Cohen's *d* für metrische Daten vergleichbar ist. Von schwachen Effekten kann bei einem $\delta \geq .147$ ausgegangen werden, von mittleren Effekten ab $\delta \geq .330$, von starken Effekten ab $\delta \geq .474$ (Cliff, 1993; Driemeyer, Spehr, Yoon, Richter-Appelt & Briken, 2013; Kromrey & Hogarty, 1998; Macbeth et al., 2011).

Hinsichtlich einer möglichen Verzerrung der Stilurteile aufgrund der sprachlichen Richtigkeit zeigen sich für alle drei Textsorten bedeutsame Effekte. Die mittleren Abweichungen zwischen korrigierten und fehlerbelassenen Textvarianten betragen für die narrativen Texte 0.15 Stufen (p = .002; δ = .159), für die argumentierenden Texte 0.25 Stufen (p < .001; δ = .317) und für die informierenden Texte 0.24 Stufen (p < .001; δ = .269). Abbildung 8.4.2.8 illustriert die relative Verteilung der Textpaare gemäß der Bewertung der beiden Textvarianten. Auch hierbei muss wiederum davon ausgegangen werden, dass die positiven Abweichungen (*korrigierte Texte besser*) bis zu einer Stufe konfundiert sind mit üblichen Kodiererschwankungen, wiederum können jedoch die negativen Abweichungen (*fehlerhafte Texte besser*) als Vergleichswert herangezogen werden, um den Anteil der Schwankungen vom Anteil des Effektes, der auf die Fehlerkorrektur zurückzuführen ist, zu trennen.

Abbildung 8.4.2.8: ***Relative Verteilung der Textpaare gemäß ihrer stilistischen Beurteilung.***



Es zeigen sich erneut keine textmusterspezifischen Unterschiede der Verteilungen (p = .222; $\delta_{arg./inf.}$ = .038; $\delta_{arg./nar.}$ = .110; $\delta_{inf./nar.}$ = .072).

Somit kann festgehalten werden, dass die Beurteilung der inhaltlichen Qualität der Schülertexte unbeeinflusst durch die sprachliche Richtigkeit der Texte erfolgt; die Beurteilung der stilistischen Qualität hingegen ist zu einem gewissen Maße durch die sprachliche Richtigkeit verzerrt.

Für das Vorliegen einer Verzerrung der Stilurteile, jedoch nicht der Inhaltsurteile, können mehrere Gründe erwogen werden. Betrachtet man die beurteilten Aspekte der einzelnen Dimensionen (vgl. Tabelle 8.4.2.1), zeigt sich, dass es, insofern man sich an

Oberflächenmerkmalen wie sprachlichen Kategorien oder Textbausteinen orientiert, Überlappungen zwischen den einzelnen Bereichen gibt. So ist beispielsweise die korrekte Tempusbildung (*ich dachte* und nicht *ich denkte*) ein zu beurteilender Aspekt der sprachlichen Richtigkeit, der textsortenadäquate Tempusgebrauch (z. B. Präsens in einer Gebrauchsanleitung) ist dem Stil zuzuordnen. Derartige Überlappungen finden sich zwischen den Dimensionen *sprachliche Richtigkeit* und *Stil* sowie *Stil* und *Inhalt*, nicht jedoch zwischen s*prachliche Richtigkeit* und *Inhalt.*

Tabelle 8.4.2.1: ***Beurteilungsrelevante Aspekte und ihre Zuordnung zu den Schreibkompetenzdimensionen.***

Sprachliche Richtigkeit	***Stil***	***Inhalt***
Orthografie Zeichensetzung		
Richtiger Gebrauch von Funktionswörtern (Präpositionen, Konjunktionen, Artikel)	Angemessener Gebrauch von Funktionswörtern	
Korrekte Tempusbildung	Textsortenadäquater und einheitlicher Tempusgebrauch	
Korrekter Satzbau	Variabilität / Textsortenkonformität des Satzbaus	
Lexikalische Korrektheit°	Lexikalische Angemessenheit und Variation	
	Elemente vorhanden? (z. B. Überschrift)	Inhaltliche Güte der Elemente
	Struktur, Gliederung / Strukturstiftung (Kohärenz)	Arrangement, Ordnung der inhaltlichen Elemente
	Einheitliche und zutreffende Perspektive und Adressierung	
		Diverse aufgabenspezifische inhaltliches Aspekte (z. B. Kontraargumente)

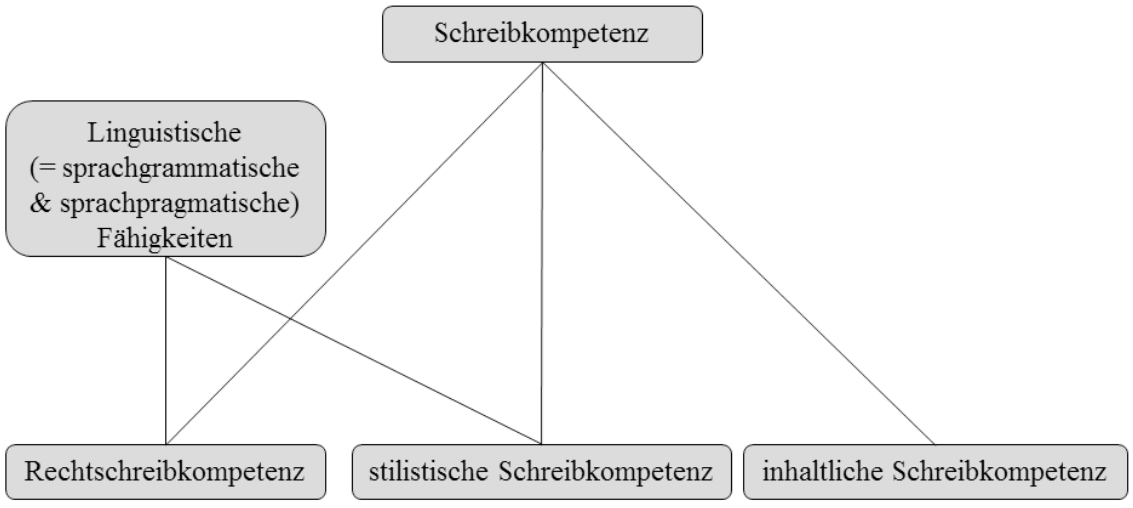
° Im Rahmen der holistischen Kodierung wurde dieser Aspekt nicht berücksichtigt.[61]

[61] Eine Unterscheidung zwischen lexikalischer Korrektheit und lexikalischer Angemessenheit erweist sich als vage und wenig trennscharf, weshalb bereits bei der Skalenentwicklung auf das Konzept der lexikalischen Korrektheit verzichtet wurde.

Auch wenn durch die mehrfache Schulung der Kodierenden sichergestellt wurde, dass diesen die Unterscheidung zwischen den Ebenen, also etwa zwischen Korrektheit und Angemessenheit von Satzbau, Tempora oder Funktionswörtern, explizit klar war, so ist dennoch zu vermuten, dass im Zuge der Beurteilung implizite Verarbeitungsprozesse und Orientierungen an diesen Oberflächenmerkmalen, also etwa an *Syntax* oder *Tempusgebrauch* (unabhängig von Korrektheit / Angemessenheit) zeitweise zutage traten. Dies untermauert die Annahme eines per definitionem impliziten Halo-Effektes.

Ebenfalls auf die Gemeinsamkeit auf sprachlicher Ebene zwischen stilistischen Aspekten und Aspekten der sprachlichen Richtigkeit bezieht sich eine zweite Erklärungsmöglichkeit bezüglich der Einflüsse auf stilistischer, jedoch nicht auf inhaltlicher Ebene. Diese Erklärungsmöglichkeit steht im Einklang mit oben dargestellten Ergebnissen und Interpretationen der Studien von Kreiner et al. (2002), Varnhagen (2000) und Figueredo und Varnhagen (2005) und dem daraus abgeleiteten Modell bezüglich der Annahmen über den Zusammenhang zwischen Personenfähigkeiten. Diesem Modell zufolge ist es möglich, dass die Beurteilenden eine gemeinsame zugrunde liegende Fähigkeit für stilistische und Rechtschreibfähigkeiten annehmen, die jedoch nicht der inhaltlichen Schreibfähigkeit zugrunde liegt (vgl. Abbildung 8.4.2.9).

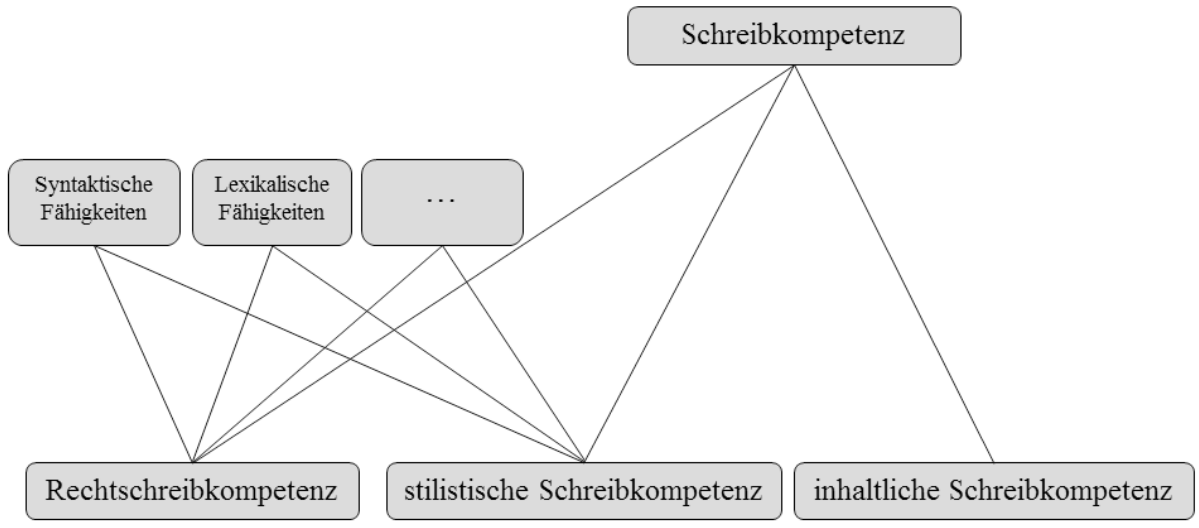
Abbildung 8.4.2.9: ***Mögliche Annahme über die sprachliche Fähigkeitsstruktur bei der Beurteilung von stilistischen, inhaltlichen und Rechtschreibfähigkeiten (1).***



Gemäß diesem Modell gingen die Beurteilenden von einem Fähigkeitskonzept aus (hier: *linguistische Fähigkeiten*), welche der *Rechtschreibkompeten*z und der *stilistischen Schreibkompetenz* zugrunde liegt, nicht jedoch der *inhaltlichen Schreibkompetenz*. Wenn dieses Konzept (*linguistische Fähigkeiten*) (gemessen an der Hierarchietiefe) weniger tief eingebettet ist als das Konzept der Schreibkompetenz, erklärt dies eine Beeinflussung der Stilurteile, nicht aber der Inhaltsurteile anhand der sprachlichen Richtigkeit. Im Detail liefe der (implizite) Beurteilungsprozess wie folgt ab: Aufgrund der sprachlichen Richtigkeit der Texte fällen die Kodierenden implizit Urteile über die Rechtschreibkompetenz. Dadurch werden implizit auch Urteile über zugrunde liegende Fähigkeiten getroffen. Wie stark die entsprechenden Einflüsse dabei sind, hängt von der konzeptuellen Nähe/Distanz ab. Ist eine dieser Fähigkeiten erst mal hinreichend (implizit) mitbeurteilt, wirkt diese Beurteilung auch auf die Beurteilung anderer Teilfähigkeiten dieses Konstrukts, in diesem Fall die Beurteilung stilistischer Schreibfähigkeiten.

Diese zweite Erklärungsmöglichkeit steht nicht in Konkurrenz zur ersten, sondern kann vielmehr auch als mögliche Ergänzung betrachtet werden, insofern, dass es sich bei der angenommen zugrunde liegenden Fähigkeit auch um mehrere angenommene Fähigkeiten handeln kann, deren Beurteilung sich an Oberflächenmerkmalen orientiert (vgl. Abbildung 8.4.2.10).

Abbildung 8.4.2.10: ***Mögliche Annahme über die sprachliche Fähigkeitsstruktur bei der Beurteilung von stilistischen, inhaltlichen und Rechtschreibfähigkeiten (2).***



Eine dritte Erklärungsmöglichkeit für den Einfluss der sprachlichen Richtigkeit auf die Beurteilung der stilistischen, nicht jedoch der inhaltlichen Qualität der Schülertexte, basiert ebenfalls auf der Oberflächenparallelität stil- und rechtschreibrelevanter sprachlicher Aspekte. Liegt eine Verletzung der sprachlichen Richtigkeit hinsichtlich eines dieser Oberflächenmerkmale vor, muss der entsprechende Fehler zunächst korrigiert werden, um die stilistische Adäquatheit beurteilen zu können. Auch wenn dies in der Mehrheit der Fälle unproblematisch sein sollte, so kann im Einzelfall eine Korrekturentscheidung in Zweifelsfällen die Entscheidung pro oder kontra stilistische Angemessenheit mitbedingen. Schreibt beispielsweise ein Schüler in einem Bericht *Zuvor hatt er die Wohnungstür abgeschlossen*, ist es entscheidend, ob der stilistische Beurteiler *hatt* zu *hatte* und somit zu einer Form des Plusquamperfekts, der berichtsadäquaten Zeitform für Vorzeitigkeit, oder zu *hat* und somit einer berichtsinadäquaten Perfektform verbessert.

Anhand der bisherigen Befundlage ist nicht zu entscheiden, welche der drei Erklärungsmöglichkeiten die zutreffende ist. Allerdings sind die drei unterschiedlichen Erklärungen mit verschiedenen Implikationen für die untersuchten Aspekte in den folgenden Untersuchungsteilen verbunden:

Erklärungsmöglichkeit 1 nimmt an, dass trotz der Fähigkeit der Kodierenden, im expliziten Diskurs zwischen Korrektheit und Angemessenheit verschiedenerer sprachlicher Oberflächenelemente unterscheiden zu können, im Kodierprozess diese Unterscheidungsfähigkeit nicht immer zum Einsatz kommt und sich implizit nur an den Oberflächenmerkmalen orientiert wird. In diesem Fall sollte eine Beeinflussung sowohl unter Fehlerpräsenz als auch -absenz vorliegen. Einflussreiche Fehlertypen sollten ausschließlich solche sein, welche eine sprachliche Klassifikationsebene betreffen, die sowohl orthografisch-grammatische als auch stilistische Aspekte umfasst, d. h. nur grammatische Fehler; orthografische Fehler und Zeichensetzungsfehler sollten keinen Einfluss zeigen.

Erklärungsmöglichkeit 2 geht davon aus, dass die Kodierenden eine gemeinsame zugrunde liegende Fähigkeit bzw. mehrere gemeinsame Fähigkeiten für Rechtschreibkompetenzen und stilistische Schreibkompetenzen annehmen, die jedoch nicht den inhaltlichen Schreibfähigkeiten zugrunde liegt. Aufgrund dieser implizit angenommenen zugrunde liegenden Fähigkeit(en) beeinflusst die sprachliche Richtigkeit unbewusst (im Sinne eines Halo-Effekts) das stilistische Urteil. Auch gemäß dieser Möglichkeit wird eine Beeinflussung sowohl unter Fehlerpräsenz als auch unter Fehlerabsenz erwartet. Unter der vorgestellten Modellierung, welche die Rechtschreibleistung vollständig (inklusive aller Rechtschreibfehler) als Ausdruck

der *linguistischen Kompetenz* darstellt (Abbildung 8.4.2.9), sollten sich keine Unterschiede hinsichtlich verschiedener Fehlertypen zeigen. Unter der Modellierung, welche mehrere feingliedrigere gemeinsame zugrunde liegende Fähigkeiten annimmt (Abbildung 8.4.2.10), sollten nur bestimmte Fehlertypen (analog zu Erklärungsmöglichkeit 1) verzerrungs-evozierend sein.

Erklärungsmöglichkeit 3 schließlich nimmt an, dass notwendige Korrekturentscheidungen die stilistischen Entscheidungen mitbeeinflussen. Dieser Erklärungsansatz basiert nicht auf unbewussten Urteilsbeeinflussungen, vermeintliche Verzerrungseffekte wären in diesem Falle nicht als Halo-Effekte zu interpretieren. Gemäß diesem Ansatz würde eine Beeinflussung nur unter Fehlerpräsenz bestehen. Einflussreiche Fehlertypen sollten solche sein, die sowohl auf orthografisch-grammatischer als auch auf stilistischer Ebene vertreten sind bzw. die die jeweils andere Ebene in der Interpretation beeinflussen können, d. h. grammatische Fehler und syntaktisch relevante Zeichensetzungsfehler, i. e. Satzkommata und Satzschlusszeichen.

8.5. Untersuchungsteil II: Verzerrungen unter Fehlerpräsenz und -absenz oder ausschließlich unter Fehlerpräsenz

8.5.1. Methoden

8.5.1.1. Datenbasis

Zur Untersuchung der Fragestellung, ob eine Verzerrung der Urteile nur unter Fehlerpräsenz auftritt oder entsprechende Urteile sowohl von Fehlerhaftigkeit (negativ) als auch von Fehlerfreiheit (positiv) beeinflusst sind, wurden die 430 Schülertexte in beiden Varianten (fehlerbelassen und korrigiert) aus Untersuchungsteil I herangezogen.

8.5.1.2. Analysen

Zwei verschiedene Messmodelle wurden mit der Statistiksoftware *Mplus* (Version 5.21) berechnet und kontrastiert. In beiden Modellen wurde je ein latenter Faktor *Inhalt* und ein latenter Faktor *Stil* auf der Basis der jeweils beiden Kodiererurteile modelliert; beide Kodiererurteile wurden jeweils mit derselben Ladung einbezogen, Stufengrenzen (*thresholds*)

sowie Residualvarianzen wurden ebenfalls für beide Urteile gleichgesetzt.[62] In beide Modelle gingen alle Texte, gruppiert nach Textvariante (fehlerbelassene Texte; fehlerkorrigierte Texte), ein. In Modell I wurden die Ladungen, Stufengrenzen und Residualvarianzen zwischen den Gruppen frei geschätzt. In Modell II wurden die Ladungen, Stufengrenzen und Residualvarianzen für die beiden Textvarianten gleichgesetzt.

Modell I nimmt Unabhängigkeit zwischen den Gruppen an und modelliert somit eine Situation, in welcher die Beurteilung fehlerhafter Texte einem anderen Konstrukt entspricht als die Beurteilung weitgehend fehlerfreier Texte. Es wird gleichsam auf Basis der fehlerfreien Texte ein Konstrukt *Stil* und auf Basis der fehlerhaften Texte ein Konstrukt *Stil + Aspekte der sprachlichen Richtigkeit* modelliert.

Modell II nimmt hingegen strikte Messinvarianz an (vgl. Schroeders & Wilhelm, 2011) und modelliert eine Situation, in welcher die Beurteilung beider Textvarianten ein einheitliches Konstrukt ist. Modelliert wird also auf der Basis sowohl der fehlerfreien als auch der fehlerhaften Texte ein Konstrukt *Stil + Aspekte der sprachlichen Richtigkeit*.

Modell I indizierte somit bei hinreichend besserer Modellpassung eine Verzerrung ausschließlich unter Fehlerpräsenz, während Modell II bei hinreichend besserer Modellpassung für eine Urteilsverzerrung sowohl unter Fehlerpräsenz als auch unter Fehlerabsenz spräche. Zum Vergleich beider Modelle wurde ein χ^2-Differenzentest berechnet.

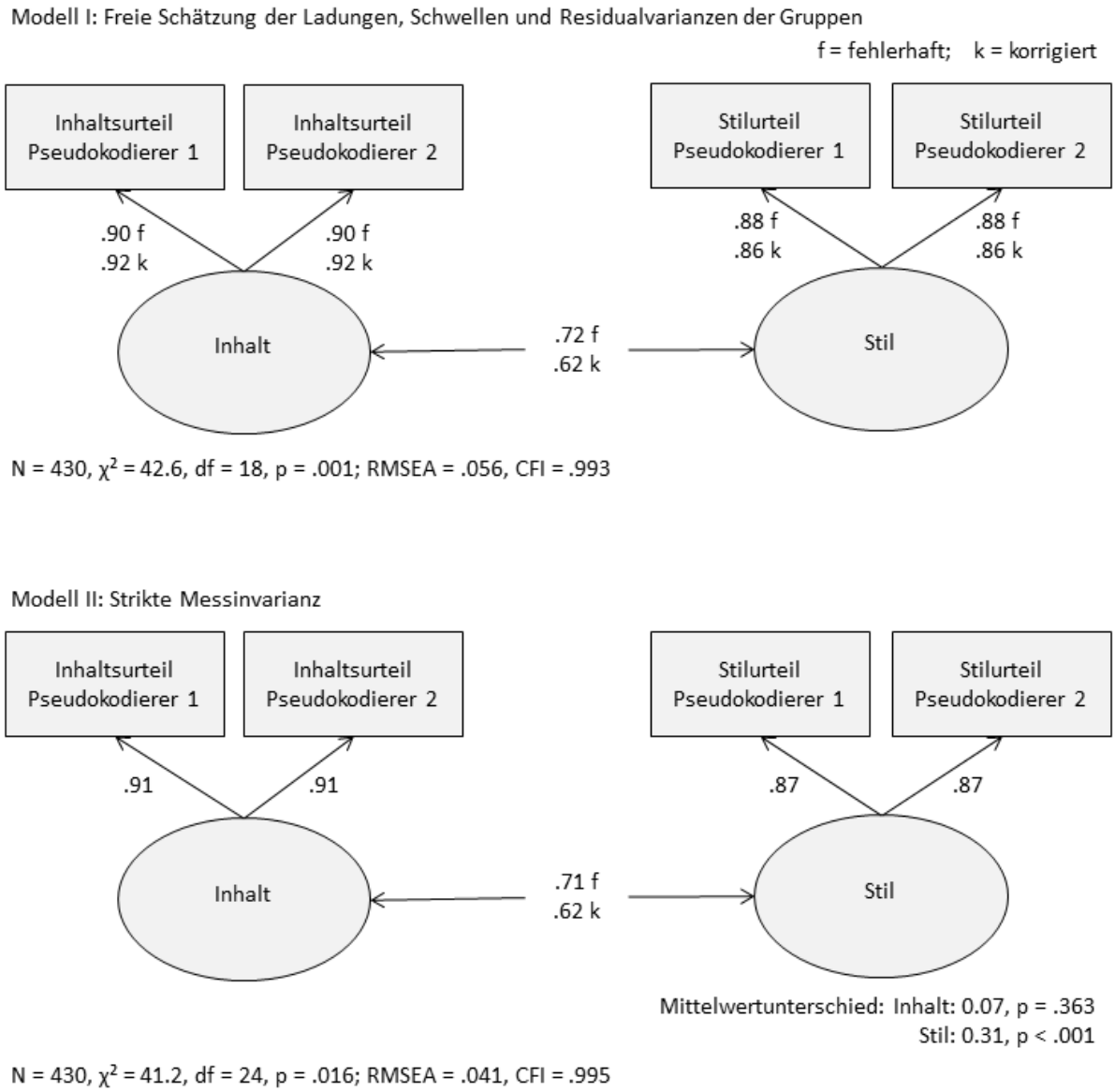
Des Weiteren wurde im messinvarianten Modell nochmals auf latenter Ebene und über Textmuster hinweg der Mittelwertsunterschied zwischen beiden Gruppen geprüft.

8.5.2. Ergebnisse und Diskussion

Zunächst zeigt sich, dass sich auch auf latenter und somit messfehlerkorrigierter Ebene die Beurteilung der inhaltlichen Qualität der Schülertexte zwischen fehlerhaften Texten und weitgehend fehlerfreien Texten nicht unterscheidet (mittlere standardisierte Abweichung: 0.07; $p = .363$). Die Beurteilung der stilistischen Qualität der Schülertexte unterscheidet sich hingegen bedeutsam; weitgehend fehlerfreie Texte werden durchschnittlich um circa ein Drittel einer Standardabweichung (0.31; $p < .001$) höher eingestuft als fehlerhafte Texte.

[62] Bei den beiden Kodierurteilen handelt es sich entsprechend des Raterdesigns um die Urteile von wechselnden Kodiererpaarungen und nicht von zwei Personen. Es handelt sich somit um Urteile zweier Pseudorater. Pseudorater werden als strukturgleich angenommen. Dieser Annahme wird mit der Gleichsetzung der Ladungen Rechnung getragen (vgl. Böhme et al., 2009).

Abbildung 8.5.2.1: ***Modellierung der latenten Faktoren Inhalt und Stil nach Gruppen (fehlerhaft vs. korrigiert) in einem frei geschätzten Modell (I) und einem strikt messinvarianten Modell (II).***



Der Vergleich der beiden Messmodelle zur Prüfung der Fragestellung, ob eine Urteilsverzerrung nur unter Fehlerpräsenz oder auch unter Fehlerabsenz vorliegt, wird in Abbildung 8.5.2.1 illustriert. Modell I weist mit einem RMSEA von .056 und einem CFI von .993[63] eine adäquate bis gute Modellpassung auf. Modell II weist mit einem RMSEA von .041 und einem CFI von .995 ebenfalls eine gute Modellpassung auf. Der Differenzentest mit $\chi^2 = 5.7$; $df = 6$;

[63] Standardmäßig wird bei einem RMSEA ≤ .05 von einer guten Modellpassung, bei einem RMSEA zwischen .05 und .08 von einer adäquaten Modellpassung gesprochen, ein CFI (Comparative Fit Index) indiziert bei Werten ≥ .95 eine gute Modellpassung (Arbuckle & Wothke, 1999; Byrne, 2001; Hu & Bentler, 1999).

p = .45 indiziert, dass sich die Modelle in ihrer Passung nicht bedeutsam voneinander unterscheiden.[64]

Allerdings unterscheidet sich das Ladungsmuster in Modell I auch nur äußerst gering von dem in Modell II. Die frei geschätzten Ladungen betragen für die Inhaltsbeurteilung .90 für die fehlerhaften, .92 für die korrigierten Texte, für die Beurteilung des Stils .88 für die fehlerhaften, .86 für die korrigierten Texte. Die mit den invariant gesetzten Ladungen (.91 für Inhalt; .87 für Stil) nahezu identischen frei geschätzten Ladungen zeigen, dass das frei geschätzte Modell ohnehin sehr ähnlich zum messinvarianten Modell generiert wurde, was für die Interpretation des Vorliegens eines einheitlichen Konstrukts, i. e. *Stil + Aspekte der sprachlichen Richtigkeit*, spricht.

Da unter Fehlerfreiheit keine negativen Verzerrungen der Stilurteile auf Basis der sprachlichen Richtigkeit möglich sind, spricht dies dafür, dass Urteilsverzerrungen aufgrund der sprachlichen Richtigkeit sowohl negativ unter Fehlerpräsenz als auch positiv unter Fehlerabsenz erfolgen. Diese Interpretation ist jedoch nur als eine mögliche und vorläufige zu betrachten. So könnte das Befundmuster auch daher rühren, dass Einflüsse der sprachlichen Richtigkeit bei der Beurteilung fehlerhafter Texte nur unter bestimmten Fehlertypen oder Fehlermengen auftreten und der Anteil solcher Texte an der Gesamtmenge der fehlerhaften Texte zu gering ist, um in einem Modellvergleich, für welchen alle fehlerhaften Texte als eine Klasse behandelt werden, unterscheidungswirksam zu werden.[65]

[64] An dieser Stelle sei angemerkt, dass die Analyse auch separat nur für die Stilurteile durchgeführt wurde, da der Verdacht bestand, dass die Miteinbeziehung der Inhaltsurteile, bei welchen keine Urteilsverzerrungen festgestellt werden konnten, die kontrastierten Modell in Richtung Gleichheit verlagern könnten. Dieser Verdacht konnte jedoch nicht bestätigt werden; der Modellvergleich ausschließlich auf Basis der Stilurteile führte zu einem dem hier vorgestellten analogen Befundmuster.

[65] Fehlertyp- und Fehlermengenabhängigkeit der Urteilsverzerrungen werden im Folgenden im Rahmen der Untersuchungsteile III und IV untersucht.

8.6. Untersuchungsteil III: Fehlertypabhängigkeit der Verzerrung unter Heranziehung der analytischen Kriterien

8.6.1. Methoden

8.6.1.1. Verwendung analytischer Variablen der sprachlichen Richtigkeit

Zur Beantwortung der Frage, ob gefundene Verzerrungen von bestimmten Fehlertypen abhängig sind, wurden im Rahmen dieses Untersuchungsteils einige der analytischen Variablen aus der Normierungsstudie herangezogen, i. e. diejenigen analytischen Variablen, welche dem Bereich *sprachliche Richtigkeit* zuzuordnen sind: *Orthografie*, *Grammatik*, *Zeichensetzung*. Untersucht wurden Zusammenhänge zwischen den entsprechenden Ausprägungen dieser analytischen Variablen und dem Abweichungsgrad zwischen fehlerhaften und fehlerkorrigierten Texten.

8.6.1.2. Datenbasis

Von den 430 Textpaaren aus den Untersuchungsteilen I und II wurden 391 Textpaare in die Analyse miteinbezogen, das sind alle Textpaare, für welche auch analytische Kodierungen vorlagen.[66]

8.6.1.3. Analysen

Als Werte für die analytischen Kriterien wurden die Mittelwerte der beiden Pseudokodierer herangezogen. Da es sich bei den analytischen Variablen um dichotome Urteile handelt, konnten die gebildeten Mittelwerte die Werte 0, 0.5 und 1 annehmen.[67] Zusammenhänge zwischen stilistischen Abweichungen (wie in Untersuchungsteil I erfasst als Differenz der textbezogenen Stilurteile für die fehlerhaften und die korrigierten Textvarianten) und den

[66] Im Rahmen der Normierungsstudie wurden aus ökonomischen Gründen nicht alle Texte anhand aller Skalen kodiert (vgl. Kapitel 3.5.). Aus diesem Grunde lagen für einen geringen Teil der in dieser Untersuchung verwendeten Texte keine analytischen Werte vor.

[67] Dieses Verfahren wurde gewählt, da im Rahmen der analytischen Kodierung bei den Variablen im Bereich *Sprachliche Richtigkeit* mit Fehlerquotienten gearbeitet wurde (bspw. 4 orthografische Fehler pro 100 Wörter oder 2 grammatikalische Fehler pro 100 Wörter). Uneinheitliche Urteile zwischen den beiden Pseudokodierern sollten in der Regel aufgrund geringer Fehler- oder Wortzählunterschiede entstanden sein. Diese Fälle sollten sich somit nahe an der Grenze des Erfülltheitskriteriums bewegen. Die Bildung der Mittelwerte wird dieser Grauzone gerecht und ordnet Texte im Schwellenbereich dem Wert 0.5 zu.

analytischen Werten wurden mittels Spearman-Rangkorrelationen ermittelt. Zur Kontrolle wurden die Analysen auch für die Inhaltsurteile durchgeführt.

8.6.2. Ergebnisse und Diskussion

Es zeigen sich erwartungsgemäß keine bedeutsamen Korrelationen zwischen den Differenzen in den Inhaltsurteilen und den analytischen Kriterien und somit keine Zusammenhänge zwischen Abweichungen in der inhaltlichen Beurteilung und der analytischen Beurteilung der sprachlichen Richtigkeit. Allerdings treten auch für den Zusammenhang zwischen den Ausprägungen der analytischen Kriterien und der stilistischen Beurteilungsabweichung keine oder nur sehr schwache Effekte auf; lediglich für *Grammatik* zeigt sich ein bedeutsamer (p = .031), aber effektschwacher (r = .11) Zusammenhang (vgl. Tabelle 8.6.2.1).

Tabelle 8.6.2.1: ***Zusammenhänge zwischen analytischen Kriterien der sprachlichen Richtigkeit und Abweichungen zwischen fehlerhaften und fehlerkorrigierten Texten hinsichtlich der inhaltlichen und stilistischen Bewertung (Werte = Spearman's ρ).***

Kriterium	inhaltliche Abweichung	stilistische Abweichung
Orthografie	- .05	-.06
Grammatik	-.07	-.11*
Zeichensetzung	-.03	-.07

* $p < .05$

Das Vorliegen keiner oder nur sehr schwacher Effekte deutet zunächst auf weitgehende Fehlertypunabhängigkeit der stilistischen Urteilsdifferenzen hin. Allerdings könnte auch die Art und Weise, wie Fehlertypen im Rahmen dieser Analyse quantifiziert worden sind, mögliche Fehlertypeffekte überdecken. So lässt sich zum einen vermuten, dass die analytischen Kriterien aufgrund der dichotomen Erfassung eines Merkmals eine zu grobe Kategorisierung in *erfüllt* und *nicht erfüllt* vornehmen und differenzierte Abstufungen im Erfülltheitsgrad fehlen oder dass die Grenzlinie zwischen *erfüllt* und *nicht erfüllt* an einer anderen Schwelle verortet ist, die entfernt ist von den relevanten Fehleranzahlen, welche Unterschiede in der Textbeurteilung von fehlerhaften und weitgehend fehlerfreien Textvarianten auslösen. Zum anderen könnten die analytischen Kriterien zu summarisch und zu

grob kategorisiert sein, um potentiell wirksame Fehlertypen hinreichend zu unterscheiden. So werden etwa Laut-Buchstaben-Zuordnungsfehler und andere Graphemfehler, Verstöße gegen die Groß-/Kleinschreibung, Verstöße gegen die Getrennt-/Zusammenschreibung sowie einige weitere Fehlertypen unter das eine Kriterium *Orthografie* subsumiert. Um sowohl eine abgestuftere, als auch feiner kategorisierte Klassifikation von Fehlertypen vorzunehmen, wurde Untersuchungsteil IV konzipiert.

8.7. Untersuchungsteil IV: Fehlertypabhängigkeit der Verzerrung unter Fehlerzählung und -typisierung

8.7.1. Methoden

8.7.1.1. Datenbasis

Für diese Untersuchung wurden zunächst die 430 Texte, die in den Untersuchungsteilen I und II verwendet wurden, herangezogen. Unter diesen wurden diejenigen ausgewählt, die gemäß den Urteilen beider Pseudokodierer keine Abweichung zwischen der stilistischen Beurteilung der fehlerbelassenen und der fehlerbehafteten Variante aufwiesen (Abweichung/Differenz = 0; n = 102), sowie all diejenigen Texte, die gemäß den Urteilen beider Pseudokodierer eine Abweichung im stilistischen Urteil zugunsten der fehlerkorrigierten Textvariante aufwiesen (Abweichung/Differenz ≥ +1; n = 61). Insgesamt gingen somit 163 Texte in die Analyse ein.

8.7.1.2. Fehlerzählung und -kategorisierung

Für diese 163 Schülertexte wurden die in der fehlerhaften Textvariante enthaltenen Fehler gezählt und kategorisiert. Die Kategorisierung erfolgte mittels eines zuvor entwickelten Fehlerkategorisierungsmanuals. Unterschieden wurden die folgenden Fehlertypen, jeweils drei bis vier Fehlertypen wurden zusätzlich zu einer Metakategorie zusammengefasst:

- *Orthografie* – Einzelkategorien: *Lautbuchstabenzuordnung/Graphemebene*; *Groß-/Kleinschreibung*; *Getrennt-/Zusammenschreibung*; *Silbentrennung*
- *Zeichensetzung* – Einzelkategorien: *Satzkommata*; *andere Kommata*; *Satzschlusszeichen*; *sonstige Zeichensetzung*
- *Grammatik (ohne Satzbau)* – Einzelkategorien: *Morphologie (ohne Flexion)*; *Flexion*; *Präpositionen & Konjunktionen*
- *Satzbau* – Einzelkategorien: *fehlende grammatische Wörter*; *fehlende lexikalische Wörter*; *Wortstellung*; *Satzbau, massiv*[68]

Außerdem ergänzend:

- *Grammatik (total)*; setzt sich aus den Metakategorien *Grammatik (ohne Satzbau)* und *Satzbau* zusammen.

8.7.1.3. Analysen

Die Merkmale, die für die Auswahl der Texte relevant waren (Abweichung = 0 / Abweichung ≥ +1) definierten zwei Textgruppen, die Gruppe der (positiven) Abweichler und die Gruppe der Nichtabweichler. Diese Gruppen wurden hinsichtlich der Fehleranzahl in einer bestimmten Fehlerkategorie und bezüglich der Gesamtfehlerzahl verglichen. Diese Vergleiche wurden mittels Mann-Whitney-U-Tests (Mann & Whitney, 1947) vorgenommen, da für keine der Fehlertypvariablen Normalverteilung vorlag. Zur Testung auf Normalverteilung wurden Shapiro-Wilk-Tests (Shapiro & Wilk, 1965) berechnet (alle $p < .001$).

[68] Unter *Satzbau, massiv* wurden syntaktische Verletzungen gezählt, die in mehrerlei Hinsicht syntaktische Prinzipien verletzten und keine eindeutige Reparatur bestimmbar gewesen wäre, sodass auch eine einfache Kategorisierung des syntaktischen Fehlers nicht möglich war.

8.7.2. Ergebnisse

Der Vergleich der beiden Gruppen (keine Abweichung vs. Abweichung) bezüglich der Fehleranzahl bestimmter Fehlertypen in der fehlerhaften Textvariante weist bedeutsame Effekte für *Satzkommata* und die Metakategorien *Zeichensetzung* und *Grammatik (total)* aus sowie Tendenzen für *Satzschlusszeichen*, *Flexion* und *Präpositionen & Konjunktionen* sowie die Metakategorien *Grammatik (ohne Satzbau)* und die Fehleranzahl insgesamt (vgl. Tabelle 8.7.2.1).

Die spezifischen Fehlerkategorien *Silbentrennung* und *Satzbau, massiv* erweisen sich als zu niedrig besetzt, um die Ergebnisse sinnvoll interpretieren zu können. Für die Kategorie *fehlende lexikalische Wörter* wurden gemäß Korrekturmanual keine Verbesserungen vorgenommen, weshalb hier keine Unterschiede erwartbar waren, was sich in den Ergebnissen widerspiegelt.

Die Beeinflussung der Stilurteile erweist sich somit als fehlertypabhängig und tritt vor allem unter Prä-/Absenz von grammatischen Fehlern sowie Zeichensetzungsfehlern zutage. Unter den Zeichensetzungsfehlern erweisen sich auf spezifischerer Ebene vor allem syntax-indizierende Satzzeichen wie Satzschlusszeichen und Satzkommata als relevant. Auf grammatischer Ebene sind morphosyntaktische und syntaktische Fehler verzerrungswirksam. Auch die Fehleranzahl insgesamt erwies sich tendenziell als verzerrungsvergrößernd.

Tabelle 8.7.2.1: ***Vergleich von Texten mit stilistischer Beurteilungsabweichung und ohne stilistische Beurteilungsabweichung in Abhängigkeit von der Fehleranzahl bestimmter Fehlertypen.***

Fehlertyp	**keine Abweichung (mittlere Fehlerzahl)**	**Abweichung (mittlere Fehlerzahl)**	***p*-Wert (U-Test)**
1. Lautbuchstaben-/Graphemebene	6.10	7.07	.73
2. Groß-/Kleinschreibung	7.34	8.21	.81
3. Getrennt-/Zusammenschreibung	1.84	2.13	.11
4. Silbentrennung	0.01	0.00	.44
5. Satzkommata	4.64	6.46	.04*
6. andere Kommata	2.02	2.51	.35
7. Satzschlusszeichen	1.02	1.79	.07^{+}
8. sonstige Zeichensetzung	0.35	0.48	.97
9. Morphologie (ohne Flexion)	0.59	0.36	.36
10. Flexion	2.02	3.26	.09^{+}
11. Präpositionen/Konjunktionen	1.14	1.48	.08^{+}
12. fehlende lexikalische Wörter	0.25	0.23	.76
13. fehlende grammatische Wörter	0.72	0.89	.90
14. Wortstellung	0.33	0.51	.23
15. Satzbau, massiv	0.07	0.02	.14
Metakategorien			
Orthografie (1.–4.)	15.29	17.41	.45
Zeichensetzung (5.–8.)	8.03	11.23	.01*
Grammatik ohne Satzbau (9.–11.)	3.75	5.10	.09^{+}
Satzbau (12.–15.)	1.37	1.64	.76
Grammatik total (9.–15.)	5.12	6.74	.04*
Fehleranzahl insgesamt	27.07	34.02	.09^{+}

$^{+}$ $p < .10$ * $p < .05$

8.8. Untersuchungsteil V: Abhängigkeit der Verzerrung von Textlänge und Textkomplexität

8.8.1. Methoden

8.8.1.1. Datenbasis

Zur Überprüfung, ob die gefundenen Verzerrungen von Textlänge und/oder der Textkomplexität abhängen, wurden alle Texte, die bereits Untersuchungsteil I und II zugrunde lagen, herangezogen (n = 430). Es wurde wiederum (wie in Untersuchungsteil III) neben der stilistischen Abweichung auch die inhaltliche Abweichung (zur Kontrolle) in die Analyse miteinbezogen.

8.8.1.2. Bestimmung der Längen- und Komplexitätsmaße

Bezüglich der möglichen Einflussmodifikation aufgrund von Textlänge und/oder Textkomplexität wurden verschiedene Maße für die Texte bestimmt, indem Zeichen, Wörter, Teilsätze und Sätze[69] für die Texte gezählt und im Anschluss Quotientenmaße bestimmt wurden:

Länge:

- Zeichenanzahl
- Wortanzahl
- Anzahl an Teilsätzen
- Anzahl an Sätzen

Komplexität:

- Zeichen pro Wort
- Wörter pro Teilsatz
- Wörter pro Satz
- Teilsätze pro Satz

[69] Mit *Teilsätzen* sind hier sprachliche Einheiten gemeint, die aus einem Verb und seinen Ergänzungen sowie ggf. freien Angaben (z. B. Adverbialen) bestehen. Unter *Sätze* verstehen sich sprachliche Gebilde, die in schriftlicher Form mit einem Satzschlusszeichen beendet werden. Beispielsweise bestünde der folgende (eine) Satz aus drei Teilsätzen: *Da Herr Meyer am Dienstagabend bei einem Geschäftsessen war, kann er nicht bezeugen, ob seine Frau zur vermeintlichen Tatzeit zu Hause war.* Im Deutschen existiert keine einheitliche Begriffsverwendung, die zwischen diesen beiden Satzebenen unterscheidet. Im Englischen entspräche dem hier verwendeten Ausdruck *Teilsatz* der Ausdruck *clause*, dem Ausdruck *Satz* der Ausdruck *sentence*. Auch im Deutschen findet sich bisweilen der Terminus *Sentenz* für *Satz*.

8.8.1.3. Analysen

Mittels Spearman-Rangkorrelationen wurde geprüft, ob sich bedeutsame Zusammenhänge zwischen dem Ausmaß der Abweichung in der stilistischen bzw. inhaltlichen Beurteilung fehlerhafter und fehlerkorrigierter Texte und der Textlänge bzw. Textkomplexität zeigen.

8.8.2. Ergebnisse

Tabelle 8.8.2.1: ***Zusammenhänge zwischen Textlänge bzw. Textkomplexität und inhaltlicher bzw. stilistischer Abweichung (Werte = Spearman's ρ).***

Textmerkmal	inhaltliche Abweichung	stilistische Abweichung
Textlänge		
Zeichenanzahl	-.08	-.11*
Wortanzahl	-.08	-.11*
Teilsatzanzahl	-.06	-.06
Satzanzahl	-.07	-.14**
Textkomplexität		
Zeichen/Wort	.02	.07
Wörter/Teilsatz	-.01	-.09
Wörter/Satz	-.01	.08
Teilsätze/Satz	-.01	.12*

* $p \leq .05$ ** $p \leq .01$

Wie Tabelle 8.8.2.1 darlegt, zeigen sich erwartungsgemäß keine bedeutsamen Zusammenhänge zwischen dem Grad der inhaltlichen Abweichung und der Textlänge sowie der Textkomplexität der zugrunde liegenden Texte. Für den Zusammenhang zwischen stilistischer Abweichung und Textlänge zeigen sich effektschwache, aber dennoch statistisch bedeutsame Zusammenhänge bezüglich der Zeichenanzahl, der Wortanzahl sowie der Anzahl der Sätze. D. h. bei längeren Texten kommt es zu einer geringeren Verzerrung des Stilurteils in Abhängigkeit von der sprachlichen Richtigkeit der Texte. Für den Zusammenhang zwischen stilistischer Abweichung und Textkomplexität zeigt sich lediglich ein statistisch bedeutsamer, aber wiederum effektschwacher Zusammenhang bezüglich der Anzahl der

Teilsätze pro Satz. D. h. Texte, in denen mehr Satzreihen und Satzgefüge vorkommen, führen in der Regel zu einem leicht höheren Verzerrungsgrad des stilistischen Urteils.

8.9. Zusammenfassung und Gesamtdiskussion

In Untersuchungsteil I wurde gezeigt, dass die von geschulten Kodiererinnen und Kodieren gefällten Urteile hinsichtlich der stilistischen Beurteilung von Schüleraufsätzen bedeutsam durch die sprachliche Richtigkeit der Texte beeinflusst sind. Die inhaltliche Aufsatzbeurteilung erweist sich dagegen unbeeinflusst durch die sprachliche Richtigkeit der Aufsätze. Es zeigten sich hierbei keine textmusterspezifischen Unterschiede.

Die Hypothese, dass sich im Rahmen einer Kodierung durch geschulte Raterinnen und Rater keine Verzerrungseffekte der sprachlichen Richtigkeit auf die inhaltliche Textbeurteilung zeigen sollten, hat sich bestätigt. Eine von der sprachlichen Richtigkeit unbeeinflusste inhaltliche Textbewertung erweist sich somit als möglich und praktikabel.

Eine unbeeinflusste stilistische Textbewertung scheint jedoch auch bei geschulten Urteilern nicht vollkommen zu gelingen. Die stilistischen Urteile erweisen sich als durch die sprachliche Richtigkeit der zugrunde liegenden Texte beeinflusst.

Für die Tatsache, dass sich stilistische, jedoch keine inhaltlichen Urteilsverzerrungen fanden, wurden, wie oben dargestellt, drei mögliche Erklärungen angeführt:

1. Die Kodierenden orientieren sich an sprachlichen Oberflächenmerkmalen und unterscheiden nicht hinreichend zwischen orthografisch-grammatischen und stilistischen Aspekten.

2. a. Die Kodierenden nehmen (zumindest implizit) eine zugrunde liegende Fähigkeit für stilistische und orthografisch-grammatische Fähigkeiten an.

   b. Die Kodierenden nehmen (zumindest implizit) mehrere zugrunde liegende Fähigkeiten für stilistische und orthografisch-grammatische Fähigkeiten an.

3. Oftmals beeinflussen nötige Korrekturen auf orthografisch-grammatischer Ebene stilistische Textaspekte mit.

Gemäß 1. und 2. (a. und b.) sollten Verzerrungen sowohl unter Fehlerpräsenz als auch unter Fehlerabsenz auftreten, gemäß 3. lediglich unter Fehlerpräsenz. Für eine Fehlertyp-

abhängigkeit impliziert 1. sowie 2. b., dass Verzerrungen nur für grammatische Fehler auftreten, 2. a., dass Verzerrungen für alle Fehlertypen auftreten und 3., dass Verzerrungen für grammatische Fehler und syntaktisch relevante Zeichensetzungsfehler, i. e. Satzkommata und Satzschlusszeichen, auftreten.

Die Modellierungen aus Untersuchungsteil II lieferten keine Evidenz für unabhängige Konstrukte *Stil unter Verstößen gegen die sprachliche Richtigkeit* und *Stil ohne Verstöße gegen die sprachliche Richtigkeit*; dies spricht zunächst dafür, dass die Stilurteile durchweg durch die Fehlerhaftigkeit (negativ) bzw. Fehlerfreiheit (positiv) beeinflusst sind. Der Grad der Beeinflussung beträgt circa eine Drittel Standardabweichung bzw. 1/5 Stufe, d. h. circa jeder fünfte Text wird um eine Stufe verschätzt aufgrund der sprachlichen Richtigkeit des Textes. Allerdings ist hierbei anzumerken, dass der Modellvergleich alle (weitgehend) fehlerfreien und alle fehlerhaften Texte bzw. Textvarianten kontrastiert und somit Fehlertyp-abhängigkeiten im Rahmen dieser Analyse nicht beachtet werden. So könnte das Befundmuster auch daher resultieren, dass die Einflüsse der sprachlichen Richtigkeit nur in einem Subsample der fehlerhaften Texte auftreten, dessen Umfang nicht ausreichend ist, um im übergreifenden Modellvergleich sichtbar zu werden.

Die Fehlertypenunterscheidung (Untersuchungsteil IV) zeigte bedeutsame Effekte zum einen für grammatische, vor allem satzgrammatische Fehlerkategorien, was teilweise im Einklang steht mit Befunden von Linn et al. (1972) und Marshall (1967), welche ebenfalls Einflüsse aufgrund grammatischer Fehlertypen gefunden haben. Im Einklang mit der Studie von Linn und Kollegen und entgegen den Befunden Marshalls wurden keine Effekte für rein orthografische oder auch speziell Lautbuchstabenzuordnungs- und anderen Graphemfehler gefunden. Entgegen beider Studien erwiesen sich hingegen Zeichensetzungsfehler als bedeutsam, hierunter allerdings nur die Unterkategorien *Satzschlusszeichen* und *Satzkommata*, nicht hingegen *andere Kommata* und *sonstige Zeichensetzung*. Diese Diskrepanz zwischen den Befunden hinsichtlich der Zeichensetzung rührt vermutlich daher, dass die Interpunktionsregeln im Deutschen und Englischen sich stark voneinander unterscheiden. Während im Englischen mehr (teil)satzinterne Kommaregeln vorherrschen und (teil)satztrennende Kommata meist optional sind, sind im Deutschen (teil)satzinterne Kommata selten und oftmals optional, die hochfrequenten (teil)satztrennenden Kommata obligatorisch (Braunschweig Verlag, 2014).

Bezüglich der konkurrierenden Erklärungen für das Auftreten von Verzerrungen stilistischer, nicht aber inhaltlicher Urteile, indiziert das tendenzielle Vorliegen der Verzerrungen unter

Fehlerpräsenz und unter Fehlerabsenz (pro 1. und 2.) sowie die Fehlertypabhängigkeit von grammatischen (pro 1., 2. b. und 3.) und grammatisch relevanten Zeichensetzungsfehlern (pro 3.), dass offensichtlich nicht ausschließlich eine dieser Erklärungen zutrifft, sondern mehrere bis alle der betreffenden Aspekte eine Rolle spielen. Da lediglich die dritte Erklärungsmöglichkeit auf einem Ansatz beruht, in dessen Rahmen die Verzerrungen auf andere Ursachen als Halo-Effekte zurückgeführt werden können, bleibt festzuhalten, dass die gefundenen Verzerrungen der Stilurteile aufgrund der sprachlichen Richtigkeit der zu beurteilenden Texte zumindest teilweise auf unbewusste Urteilsbeeinflussung im Sinne von Halo-Effekten zurückzuführen sind.

Dafür spricht auch, dass der Verzerrungsgrad in syntaktisch komplexeren Texten (mehr Teilsätze pro Satz) größer ist. In syntaktisch komplexeren Texten wird die Aufmerksamkeit verstärkt für das Parsing, die syntaktische Sprachverarbeitung, gebunden, der Urteilsprozess greift verstärkt auf unbewusste Verarbeitungsroutinen zurück (Frazier, 1979; Frazier & Rayner, 1982). Dass für auf kleineren sprachlichen Einheiten basierende Komplexitätsmaße (Zeichen pro Wort, Wörter pro Teilsatz, Wörter pro Satz) keine Effekte gefunden wurden, rührt vermutlich daher, dass die bewussten Verarbeitungsprozesse der Kodierenden und somit die Unterscheidungsfähigkeit von orthografisch-grammatischen und stilistischen Phänomenen im rein morphologischen und graphematischen Komplexitätsbereich (Zeichen pro Wort) sowie im syntaktisch eher mittleren Komplexitätsbereich (Wörter pro Teilsatz/Satz) prädominant sind und erst in syntaktisch komplexeren Bereichen unbewusste Verarbeitungs-prozesse aktiv werden. Dies steht auch im Einklang mit den Befunden, dass sich gerade die syntaxrelevanten Zeichensetzungsfehler als einflussreich erwiesen.

Warum der Anstieg der Textlänge keinen verzerrungsverstärkenden bzw. sogar einen leicht verzerrungsmindernden Effekt mit sich bringt, scheint zunächst unerwartet. Allerdings ist zum einen festzustellen, dass dem Faktor *Textlänge* als leseschwierigkeitsbestimmendem Merkmal vor allem im Rahmen des Zweitspracherwerbs und des frühen Schriftspracherwerbs (Grundschule / Alphabetisierung von Analphabeten) eine gewichtige Rolle zukommt (Grotlüschen & Riekmann, 2012; Jurecka, 2010; Nold & Rossa, 2007), die Stärke des Einflusses der Textlänge von Schrifttexten bei schriftspracherfahrenen Muttersprachlern ist in Frage zu stellen. Zum anderen weisen Nold und Rossa (2007) darauf hin, dass bei Kurztexten der Faktor Textlänge nicht zum Tragen kommt. Da alle 430 verwendeten Schülertexte in transkribierter Version maximal eine Seite lang waren (Maximalwerte: 320 Wörter, 54 Teilsätze, 31 Sätze), können die Schüleraufsätze durchweg als Kurztexte angesehen werden.

Weshalb sich jedoch bei den kürzeren dieser Kurztexte eine tendenziell höhere Verzerrung zeigt, bleibt dadurch unerklärt. In den Kodiererschulungen erwies es sich jedoch gerade bei kürzeren Texten als extrem schwierig, zuverlässige Stilurteile zu fällen. So lässt sich bei nur geringem vorliegendem Textmaterial beispielweise schwerer feststellen, ob die Textsorte eindeutig getroffen wurde; die Vielfältigkeit von Wortschatz und syntaktischen Konstruktionen lässt sich aufgrund der wenigen Textvorkommnisse lexikalischer und syntaktischer Einheiten kaum beurteilen. Daher könnte gerade in diesen Fällen, in welchen einige der Anhaltspunkte für die stilistische Beurteilung für die Kodierenden nicht eindeutig zu interpretieren sind, stärkere unbewusste Verarbeitungsprozesse wie solche, die Halo-Effekte begünstigen, zutage treten.

8.9.1. Implikationen für das Schreibassessment

Die Ergebnisse zeigen, dass zumindest im Rahmen von Schreibassessments, in welchen die Beurteilenden in der Unterscheidung zwischen den konzeptionell unabhängigen Ebenen *Inhalt*, *Stil* und *sprachliche Richtigkeit* geschult werden, eine von der Sprachrichtigkeit unabhängige inhaltliche Aufsatzbewertung möglich ist. Eine Beeinflussung der stilistischen Textbeurteilung durch die Sprachrichtigkeit der Texte scheint zumindest zu einem geringen Anteil in der Natur der Sache zu liegen, insofern, dass einige stilistische Beurteilungen von der Korrektur sprachlicher Fehler abhängen. Eine gänzliche Verhinderung oder Unterdrückung dieser Beeinflussung scheint somit unvermeidbar. Ob sich der Anteil der im Rahmen dieser Studie aufgezeigten durch die sprachlichen Richtigkeit induzierten Halo-Effekte bei der stilistischen Beurteilung durch weitere, intensivere Schulung vermeiden lässt, ist fraglich. Bereits im Rahmen der vorliegenden Studie mit intensiv geschulten Urteilenden zeigte sich, dass im Mittel jeder fünfte Text um eine Bewertungsstufe über- oder unterschätzt wurde.

9. Fazit und Gesamtdiskussion

In diesem abschließenden Kapitel werden die Befunde der empirischen Untersuchungen zusammengefasst und zueinander in Beziehung gesetzt. Außerdem wird der Bezug zur zentralen Frage nach der Validität der Schreibkompetenzmessung bzw. der Ergebnisse hergestellt und die praktische Relevanz der Untersuchungsergebnisse erörtert. Im Anschluss werden Einschränkungen bei der Interpretation der Ergebnisse sowie Implikationen für das Schreibassessment und den Unterricht erläutert. Abschließend wird ein Ausblick auf mögliche Anschlussforschungen im Themenfeld gegeben.

9.1. Zusammenfassung und Gesamtschau

Nach einführenden Bemerkungen zur Schreibkompetenz und deren Messung sowie zum Konzept und Begriff *Schreibkompetenz* (Kapitel 1 und 2) wurde die Studie zur Normierung von Aufgaben zur empirischen Überprüfung des Erreichens der Bildungsstandards im Kompetenzbereich *Schreiben* für das Fach *Deutsch* am Ende der Sekundarstufe I und das in diesem Rahmen angewandte Verfahren zur Messung von Schreibkompetenzen vorgestellt (Kapitel 3) sowie zentrale Ergebnisse der Normierungsstudie (Kapitel 4) präsentiert.

Forschungsziel dieser Arbeit war es, das angewandte Verfahren im Hinblick auf drei ausgewählte Validitätsaspekte zu untersuchen. Hierzu wurde im Anschluss (Kapitel 5) zunächst das Konzept *Validität* beleuchtet und der experimentaltheoretische Rahmen für die drei zentralen Forschungsstudien bzw. -teilstudien dieser Arbeit gespannt.

Die erste Teilstudie (Kapitel 6) widmete sich der Untersuchung der strukturellen Validität von *Schreibkompetenz* und der Frage, ob Schreibkompetenzen textmusterspezifisch oder textmusterunabhängig sind. Dabei wurden auch die theoretisch angenommenen Schreibkompetenzdimensionen *Inhalt*, *Stil* und *sprachliche Richtigkeit* betrachtet und zum einen das Verhältnis dieser Dimensionen zueinander und somit die interne Struktur von Schreibkompetenz(en) untersucht, zum anderen die Textmusterspezifität vs. Textmusterunabhängigkeit dieser Teildimensionen überprüft. Die Analysen basierten vorwiegend auf IRT-Modellierungen (Skalierungen) und dem Modellvergleich n-dimensionaler Modelle (wobei n gemäß den verschiedenen zu kontrastierenden theoretischen Annahmen variierte). Die Ergebnisse dieser Analysen sprechen für die Textmusterspezifität von Schreib-

kompetenzen, sprich für separate Konstrukte *narrative*, *argumentative* und *informierende Schreibkompetenz*. Für alle drei untersuchten Textmuster zeigt sich auch Textmusterspezifität von inhaltlichen und stilistischen Schreib(teil)kompetenzen. Lediglich hinsichtlich der orthografisch-grammatischen Schreib(teil)kompetenz liefern die Daten Evidenz für ein textmusterunabhängiges Konstrukt. Für alle drei Textmuster zeichnete sich darüber hinaus eine identische interne Struktur ab, mit *Inhalt+Stil* als eine und *sprachliche Richtigkeit* als zweite Dimension. Dies steht im Einklang mit Ergebnissen zu vorherigen Studien zur internen Struktur von Schreibkompetenzen (Böhme et al., 2009; Lehmann & Hartmann, 1987; A. Neumann, 2007).

Ebenso stimmen diese Befunde weitgehend mit den deskriptiven Ergebnissen der Normierungsstudie (vgl. Kapitel 4) überein; hier zeigten sich sowohl für das Gesamtkonstrukt *Schreibkompetenz* als auch für die Dimensionen der inhaltlichen und stilistischen Schreib(teil)kompetenz im Vergleich von Alters-, Geschlechts-, Schulform- und/oder Sprachhintergrundsgruppen textmusterspezifische Unterschiede, was ebenfalls für die Textmusterspezifität von Schreibkompetenzen spricht. Für die Dimension der sprachlichen Richtigkeit zeigten sich in den Gruppenvergleichen mit einer Ausnahme, welche sich jedoch auf curriculare Ursachen zurückführen lässt (vgl. Kapitel 4.4.), keine textmusterspezifischen Unterschiede.

Im Rahmen der zweiten Forschungsteilstudie (Kapitel 7) wurde der Zusammenhang zwischen Lese- und Schreibkompetenzen untersucht und der Kernfragestellung nachgegangen, in welchem Umfang bei der Messung von Schreibkompetenzen aufgrund der textuellen Präsentation der Instruktion und weiteren Stimulusmaterials Lesefähigkeiten und somit konstruktirrelevante Varianz miterfasst wurden. Dabei musste zunächst berücksichtigt werden, dass sprachliche Kompetenzen – somit auch *Lesen* und *Schreiben* – hohe Zusammenhänge untereinander aufweisen (u. a. Jude, 2008). Um dennoch den Einfluss spezifischer Lesekompetenzanteile bei der Bearbeitung der Schreibaufgaben erfassen zu können, wurden die Stimulus- und Instruktionstexte der Schreibaufgaben nach leseschwierigkeitsbestimmenden Merkmalen klassifiziert; konkret wurden somit Lesekompetenzanteile fokussiert, welche beim Lesen von Texten unterschiedlicher sprachlicher Schwierigkeit relevant sind. Im Rahmen von Mehrebenen-Moderator-Analysen wurde untersucht, ob der Zusammenhang zwischen Lese- und Schreibkompetenz über Aufgaben hinweg in Abhängigkeit dieser Merkmale variiert. Dabei zeigte sich, dass sich – wenn auch statistisch bedeutsam – lediglich 1.6 % der Gesamtvarianz in den Schreib-

leistungen auf die aufgabenspezifischen Leseanforderungen zurückführen lassen. Hinsichtlich der erfassten Merkmale zeigten sich die syntaktische Komplexität sowie die mittlere Seltenheit der Wörter mit einer gemeinsamen Varianzaufklärungsquote von 92 % als schwierigkeitsbestimmend.

Die dritte Teilstudie (Kapitel 8) widmete sich schließlich ebenfalls der möglichen Miterfassung konstruktirrelevanter Varianz. Im Rahmen der Untersuchungen wurde hierbei geprüft, ob bei der Erfassung der inhaltlichen und stilistischen Schreib(teil)kompetenzen die Urteile der Bewertenden durch die sprachliche Richtigkeit der Texte beeinflusst werden und sogenannten Halo-Effekten (Thorndike, 1920) unterliegen. Dabei zeigte sich, dass die inhaltliche Beurteilung weitestgehend unabhängig von orthografisch-grammatischen Textmerkmalen stattfindet, die stilistische Beurteilung jedoch zu einem gewissen Teil (circa ein Fünftel aller Texte) durch die sprachliche Richtigkeit der Texte verzerrt wird. Zusatzanalysen lieferten Evidenz dafür, dass diese Verzerrung tendenziell nicht nur negativ (unter Fehlerpräsenz), sondern auch positiv (unter Fehlerabsenz) erfolgt. Darüber hinaus zeigten ergänzende Analysen, dass diese Verzerrungen fehlertypabhängig sind und vor allem bei grammatischen Fehlern und syntaktisch relevanten Zeichensetzungsfehlern auftreten, außerdem verstärkt bei syntaktisch komplexeren Texten. Keine Abhängigkeiten zeigten sich für Textmuster, Textlänge und andere als die genannten Fehlertypen wie orthografische Fehler oder syntaktisch unbedeutende Zeichensetzungsfehler.

Sowohl die Ergebnisse zur Untersuchung der Fehlertyp- und Textkomplexitätsabhängigkeit der Urteilsverzerrungen, als auch die Ergebnisse der zweiten Forschungsteilstudie (Kapitel 7), die zeigten, dass bei syntaktisch komplexeren Instruktions- und Stimulustexten der Einfluss der Lesekompetenz auf die gezeigten Schreibleistungen steigt, stehen im Einklang mit psycholinguistischen Untersuchungen zur Sprachverarbeitung und untermauern die Relevanz der Syntax beim Verstehen von Texten und bei der Sprachrezeption im Allgemeinen (Frazier, 1979; Frazier & Rayner, 1982). Der Einfluss der syntaktischen Komplexität zeigt sich hierbei sowohl auf Ebene der Textschreiber (beim Lesen des Stimulustextes) als auch auf Ebene der Textbeurteiler (beim Lesen und Bewerten des Schreibprodukts).

9.2. Bezug zur Validität und praktische Relevanz

Die übergeordnete Fragestellung dieser Arbeit bzw. des Forschungsteils dieser Arbeit lautete: „Wie valide ist das vorgestellte Schreibkompetenzmessungsverfahren?“ / „Wie valide sind die Ergebnisse des vorgestellten Schreibkompetenzmessungsverfahrens?“ / „Wie valide sind die Interpretationen der Ergebnisse des Schreibkompetenzmessungsverfahrens?“[70]

Vollumfänglich lässt sich diese Frage natürlich nicht im Rahmen der vorliegenden Arbeit beantworten, jedoch konnten drei Aspekte dieser Frage untersucht werden:

1. Ist eine textmusterspezifische Erfassung und Modellierung von *Schreibkompetenz* oder eine textmusterunabhängige Erfassung und Modellierung die geeignete? Bezogen auf das Vorgehen in der Normierungsstudie und bei der Genese der Kompetenzstufenmodelle: Wird die dimensionale (textmusterspezifische) Erfassung von Schreibkompetenzen der Struktur des psychologischen Konstrukts *Schreibkompetenz* gerecht?
2. Inwiefern werden Schreibkompetenzen mit diesem Verfahren unabhängig von schreibkompetenzirrelevanten Lesefähigkeiten erfasst? Wird bei der Schreibkompetenzmessung konstruktirrelevante Varianz, die auf die Lesekompetenz der Scheiber zurückzuführen ist, miterhoben?
3. Gelingt die Erfassung von inhaltlichen und stilistischen Schreibkompetenzen unabhängig von Aspekten der sprachlichen Richtigkeit? Oder wird bei der Messung von inhaltlichen und stilistischen Schreibkompetenzen im Rahmen der Beurteilung konstruktirrelevante Varianz, welche auf die sprachliche Richtigkeit der Texte zurückzuführen ist, miterhoben?

Hinsichtlich der ersten Fragestellung lässt sich festhalten, dass eine textmusterspezifische Modellierung von Schreibkompetenz mit einer besseren Modellpassung verbunden ist als eine textmusterunabhängige. Dies trifft auch auf die stilistische und inhaltliche Schreib(teil)-kompetenz zu, jedoch nicht auf die Dimension der sprachlichen Richtigkeit.

Im Rahmen des hier vorgestellten der Normierungsstudie zugrunde liegenden Verfahrens wurden textmuster- und teilweise sogar aufgabenspezifische Global-, Inhalts- und Stilskalen verwendet sowie textmusterspezifische Kompetenzstufenmodelle generiert; die eingesetzte Skala zur Erfassung der sprachlichen Richtigkeit hingegen war aufgaben- und textmuster-übergreifend konstant (vgl. Kapitel 3.3. sowie die zugehörigen Anhänge A.3.3.1 bis

[70] Zur Diskussion auf welcher Ebene (Verfahren, Ergebnisse oder Ergebnisinterpretation) der Validitätsbegriff anzuwenden ist, vgl. Kapitel 5.1.

A.3.3.11). Die Erfassung der Kompetenzen und Teilkompetenzen erfolgte somit in Übereinstimmung mit der textmusterspezifischen Struktur von Schreibkompetenzen und kann somit nach derzeitigem Kenntnisstand als strukturell valide angesehen werden.

Bezüglich der zweiten Fragestellung ist festzustellen, dass die Analysen zum Einfluss der Lesekompetenz auf die gemessenen Schreibleistungswerte zwar einen Effekt von 1.6 % als statistisch bedeutsam ausweisen, dieser Effekt von unter 2 % kann jedoch als praktisch irrelevant angesehen werden. Dies gilt vor allem in Anbetracht dessen, dass ein gänzlicher Verzicht auf den Einsatz von Stimulus- und Instruktionstexten für Textproduktionsaufgaben faktisch nicht möglich ist, da ein solcher Einsatz für eine Engführung der Aufgabe sorgt und eine standardisierte Auswertung somit erst ermöglicht wird. Darüber hinaus besteht zwar die prinzipielle Möglichkeit, dass unter Einsatz von Aufgaben, die eine höhere Anforderung an die Lesekompetenz stellen als die verwendeten, der Effekt größer ausfallen könnte, faktisch werden jedoch im Rahmen von Schulleistungsstudien nach heutigen Standards die eingesetzten Aufgaben stets von Personen (Lehrkräften, Fachdidaktikern etc.) entwickelt, die eine solch gesteigerte Aufgabenschwierigkeit bewusst vermeiden.

Mit Blick auf die dritte Fragestellung lässt sich festhalten, dass neben der Erfassung der inhaltlichen Schreib(teil)kompetenzen, welche im Rahmen der Beurteilung unbeeinflusst durch die sprachliche Richtigkeit der zugrunde liegende Texte erfolgen konnte, die Ermittlung stilistischer Schreib(teil)kompetenzen in einem gewissen Umfang (circa 20 % aller Texte/ Personen) durch die sprachliche Richtigkeit der zugrunde liegenden Texte beeinflusst ist. Allerdings muss – wie in Kapitel 8 ausführlicher dargelegt – dabei beachtet werden, dass sich vermutlich nur ein Teil, wenn auch der größere, dieser Einflüsse auf Halo-Effekte zurückführen lässt, ein anderer Teil auf fehlerhafte sprachliche Strukturen zurückzuführen ist, welche man auf verschiedene Arten korrigieren kann, auf eine stilistische und auf eine orthografisch-grammatische. Darüber hinaus ist der Grad der Verzerrung (überwiegend eine Skalenstufe auf einer vierstufigen Skala) eher gering und bewegt sich im Rahmen der üblichen Kodiererabweichungen zwischen Personen bei der Textbeurteilung anhand mehrstufiger Skalen (vgl. Kapitel 3.5.). Dabei unterlag gerade die stilistische Beurteilung trotz mehrfacher Schulung der Kodierenden am stärksten subjektiven Urteilsunterschieden, was sich auch in (im Vergleich zu den anderen Skalen) niedrigeren Interraterreliablitäten widerspiegelte (vgl. Tabelle 3.5.2). Aufgrund des geringen Grades der Verzerrung und ihr Auftreten in einem ohnehin stärker subjektiven Einflüssen unterliegenden Bewertungskontext

wie der stilistischen Beurteilung ist auch hier die praktische Relevanz der gefundenen Verzerrungseffekte nur als gering einzustufen.

Die Ergebnisse der Untersuchungen zur möglichen konstruktirrelevanten Varianz bei der Messung von Schreibkompetenzen erweisen sich somit praktisch als kaum relevant und somit nicht wesentlich validitätseinschränkend. Das der Normierungsstudie zugrunde liegende vorgestellte Verfahren zur Schreibkompetenzmessung (bzw. entsprechende Ergebnisse bzw. deren Interpretationen) kann (bzw. können) somit mit Blick auf die drei untersuchten Validitätsaspekte als äußerst valide angesehen werden.

9.3. Grenzen und Einschränkungen der Untersuchungen

Auch wenn die hier vorgestellten Studien in vielerlei Hinsicht hohen Standards genügen, so sind die Ergebnisse dennoch in einigen Aspekten vorsichtig zu interpretieren, vor allem insofern man sie verallgemeinern möchte. Der größte einschränkende Faktor hierbei ist die Anzahl der eingesetzten Aufgaben. So kamen in der Hauptstudie, welche in Kapitel 3 vorgestellt wurde und die Datenbasis für die Untersuchungen in den Kapiteln 6 und 8 bildete, insgesamt lediglich zwölf Aufgaben zum Einsatz, vier je Textmuster. Da die Kompetenzstufenmodelle textmusterspezifisch entwickelt wurden (vgl. Kapitel 3), bedeutet dies auch, dass die ermittelten Stufenverteilungen pro Modell (vgl. Kapitel 4.1.) lediglich auf vier Aufgaben beruhen. Auch im Rahmen der Analysen der manifesten Schreibleistungsdaten in Kapitel 6, in welchen uneinheitliche Effekte im Vergleich der Textmuster auftraten, zeigte sich, dass, insofern die Varianz innerhalb dieser vier Aufgaben größer ist, auch textmuster-spezifische Effekte weniger stark zutage treten können.

Auch erwies sich die Form der Aufgabenverknüpfung im Rahmen des Testdesigns in der Normierungsstudie für Analysen zur Textmusterspezifität als nur begrenzt geeignet. Aufgrund der Spiral-Verknüpfung (vgl. Kapitel 3.4.), welche je zwei Nachbaraufgaben in einer geordneten Aufgabenabfolge miteinander verbindet, war jede Aufgabe nur mit zwei anderen Aufgaben direkt verbunden. Zusätzlich waren die Aufgaben nach Textmustern sortiert. Diese Sortierung sorgte zwar einerseits – was positiv hervorzuheben ist – für eine multiple Aufgabenverknüpfung von Aufgaben eines Textmusters, andererseits wurde damit jedoch nur eine direkte Verbindung zwischen je zwei Textmustern anhand eines jeweils einzigen Aufgabenpaares etabliert. Damit basieren die ermittelten Zusammenhänge von

aufgabenspezifischen Schreibleistungen zweier textmusterdifferenter Aufgaben auch nur auf einem Aufgabenpaar pro Textmusterkombination, die Generalisierbarkeit der Ergebnisse ist stark eingeschränkt. Da jedoch bereits vor den hier vorgestellten Analysen zur Textmusterspezifität von Schreibkompetenzen fachdidaktische theoretische Annahmen ein Vorliegen dieser nahelegten, war die Stärkung der textmusterinternen Verbindungen unter Anbetracht der zur Verfügung stehenden Aufgaben- und Testheftmenge zweifelsfrei die bessere Wahl; mit einem anderen Verknüpfungsverfahren, für dessen Realisierung jedoch eine deutlich höhere Anzahl an Testheftvarianten hätte zur Verfügung stehen müssen, und/oder durch den Einsatz einer höheren Aufgabenmenge hätte jedoch diese Stärke beibehalten werden und dennoch die Anzahl der textmusterübergreifenden Verbindungen erhöht werden können.

Auch für die Analysen zur Lesekompetenzabhängigkeit der ermittelten Schreibleistungswerte, welche auf den Daten einer der Pilotierungen der Aufgaben beruht, ist die Aussagekraft der Ergebnisse durch die Anzahl der wenigen Aufgaben eingeschränkt; von lediglich sieben eingesetzten Aufgaben konnten nur sechs in die Analyse miteinbezogen werden. Für die im Rahmen der Teilstudie angewandten Mehrebenenmoderatoranalysen mit Schreibaufgaben auf Ebene 2 bedeutete dies auch das Vorliegen von lediglich sechs Ebene-2-Einheiten. Während die festen Effekte im Rahmen von Mehrebenenanalysen auch bei wenigen Ebene-2-Einheiten als weitgehend stabil gelten, werden die zufälligen Effekte jedoch häufig ver-, Residualvarianzen zumeist unterschätzt (Bell et al., 2010; Nezlek et al., 2006; van der Leeden & Busing, 1994), weshalb die Ergebnisse nur begrenzt verallgemeinerbar sind.

Neben diesen Einschränkungen weisen die Untersuchungen jedoch auch viele Stärken auf, so etwa die standardisierte, durch Testleiter durchgeführte Erhebung der Daten, der Einsatz von Expertinnen und Experten entwickelten, fachdidaktisch betreuten und mehrfach erprobten Aufgaben, die annähernde Repräsentativität der beteiligten Schülerschaft, die hohe Anzahl der Teilnehmer (rund 3000 bzw. 1700 Schülerinnen und Schüler), welche eine starke empirische Basis für die angewandten Analysen bot, sowie die wissenschaftlichen Standards genügenden Beurteilungen der Schülertexte. Darüber hinaus ist für die letzte Forschungsteilstudie bezüglich möglicher Halo-Effekte bei der Bewertung inhaltlicher und stilistischer Schreibkompetenzen der quasi-experimentelle Charakter, d. h. die Kontrolle der zu kontrastierenden Versuchsbedingungen, welche durch die Genese zweier Textvarianten, welche sich nur in der Präsenz/Absenz der orthografisch-grammatisch Fehler ceteris paribus unterschieden, hergestellt wurde, positiv hervorzuheben.

9.4. Implikationen für das Schreibassessment und den schulischen Unterricht

Das in der Normierungsstudie zum Einsatz gekommene Verfahren zur Schreibkompetenzmessung erweist sich im Hinblick auf im Rahmen dieser Arbeit im Vordergrund stehende Aspekte der Validität sowie auch, was eine Voraussetzung für das Vorliegen hoher Validität ist (Hussy, Schreier & Echterhoff, 2009, Lienert & Raatz, 1998), bezüglich der anderen beiden Hauptgütekriterien von Tests, Reliabilität und Objektivität (vgl. Kapitel 3.4., 3.5. & 9.3.), als qualitativ hochwertig. Das Verfahren kann somit als exemplarisches Vorbild für andere Schreibassessmentstudien dienen, ggf. unter Beachtung der Hinweise zur möglichen Verbesserung der oben unter 9.3. erwähnten Aspekte.

Dabei ist hervorzuheben, dass die hier vorgestellten Hauptanalysen (Kapitel 4) sowie die Validitätsuntersuchungen (Kapitel 6–8) nur eines der beiden in der Studie zum Einsatz gekommenen Auswertungsschemata fokussierten, das holistische Schema (vgl. Kapitel 3.3.); die analytischen Kriterien wurden lediglich im Rahmen der Untersuchung zur Fehlerabhängigkeit der möglichen Urteilsverzerrungen der inhaltlichen und stilistischen Textbewertung (Kapitel 8.7.) herangezogen. Auch für die Genese der Kompetenzstufenmodelle (Kapitel 3.8.) wurden diese lediglich als Zusatzinformation zur Detaillierung der Stufenbeschreibungen herangezogen, während die Skalierung auf den Daten der holistischen Auswertung beruhte.

Hierbei zeigte sich, dass anders als in der Diskussion darüber, welche Art der Auswertung (holistisch vs. analytisch; vgl. Kapitel 2.3.2.) zu bevorzugen sei, mehrfach berichtet, im Rahmen der Normierungsstudie die holistischen und semiholistischen Skalen eine im Vergleich zu den analytischen Kriterien ähnlich hohe bis höhere durchschnittliche Reliabilität aufwiesen (Kapitel 3.5.). Die hier vorgestellten Analysen stellen darüber hinaus den Einsatz des holistischen Schemas in Übereinstimmung mit den Ausführungen von E. M. White (1984, 1985) als äußerst valide heraus. Im Rahmen von Schreibassessments scheint der Einsatz eines solchen Systems aufgrund der zahlreichen Vorteile (vergleichbare Reliabilität, hohe Validität, geringerer Zeit- und Kostenaufwand, unproblematische Skalierbarkeit[71]) dem Einsatz eines analytischen Kodiersystems überlegen.

[71] *unproblematisch* meint in diesem Zusammenhang: ohne normative Setzungen (bspw. Festlegungen von Erfülltheitsgrenzen) und ohne Zusatzannahmen, die selbst erst geprüft werden müssten (vgl. Kapitel 2.3.2.).

Relativierend ist jedoch festzustellen, dass diese Schlussfolgerungen auf Basis des in der Normierungsstudie eingesetzten (und in diesem Rahmen auch so bezeichneten) holistischen Schemas beruhen. Dieses Schema beruht auf einer Globalskala und drei semiholistischen Subskalen. Dieses Schema wurde stets in Kombination geschult und eingesetzt. Es können somit keine Aussagen darüber getroffen werden, ob die Verwendung ausschließlich der Globalskala zu ähnlichen Ergebnissen geführt hätte. Des Weiteren zeigen die Analysen zur internen Struktur von Schreibkompetenzen, dass zumindest inhaltliche und stilistische Schreibfähigkeiten einerseits und orthografisch-grammatische Schreibfähigkeiten zwei distinkte Dimensionen sind, deren getrennte Erfassung je nach Kontext lohnenswert und sinnvoll sein kann. Hinsichtlich der semiholistischen Skalen ist allerdings anzumerken, dass, wie der Begriff *semiholistisch* bereits indiziert, diese auf mittlerer Ebene (zwischen holistischen Globalurteilen und analytischer kriterialer Erfassung) anzusiedeln sind. In anderen Kontexten werden solche Skalen auch als *analytische Skalen* und lediglich allumfassende Globalurteile als *holistisch* klassifiziert (u. a. Chita, 2008; Harsch, 2005).

Für den Schulunterricht und somit im Rahmen einer Individualdiagnostik, an welche Folgemaßnahmen hinsichtlich der Unterrichtsgestaltung sowie gezielte Fördermaßnahmen anschließen, können holistische und semiholistische Skalen jedoch zu unkonkret sein. Wenn es darum geht, gezielt die Schwächen von Schreibern im Detail aufzudecken, bieten analytische Kriterien zur Textbeurteilung die Möglichkeit, diese Schwächen gezielt zu detektieren (Beck, 1979; Hamp-Lyons, 1991; 1996; Hofen, 1980; A. Neumann, 2007; Weigle, 2002). Lediglich im Rahmen einer finalen Leistungsbeurteilung (beispielsweise im Rahmen von Abschlusstests oder für die Leistungsbewertung am Schuljahresende), böte sich auch der Einsatz holistischer Beurteilungsinstrumente an.

Generell ist zu konstatieren, dass die im Rahmen dieser Arbeit dargelegten Stärken des der Normierungsstudie zugrunde liegenden Verfahrens nicht ohne Weiteres auf schulische Beurteilungskontexte übertragbar sind. Die Bedingungen, unter welchen schulische Schreibleistungsmessungen und -beurteilungen stattfinden, unterscheiden sich nicht unerheblich von Assessmentbedingungen. So sind den Lehrkräften die Schülerinnen und Schüler vertraut, eine Leistungsbewertung erfolgt nicht anonym, was das Potential auftretender Halo-Effekte aufgrund von personenbezogenen Einstellungen und Meinungen (Ingenkamp, 1971; Schröter, 1971; Valtin, 2002) und somit validitätsmindernder Faktoren erhöht. Auch die Reliabilität von Lehrerurteilen dürfte weitaus hinter dem hier erreichten

Niveau zurückbleiben, da der Bewertung kein Training in der Textbeurteilung und kein Austausch über diese wie im Rahmen von Kodiererschulungen vorausgehen.

Dennoch stellen die Analysen dieser Arbeit auch Befunde bereit, welche als unterrichtsrelevant interpretiert und genutzt werden können, so etwa das Ergebnis, dass bei der Bearbeitung von Schreibaufgaben kaum praktisch bedeutsame Einflüsse schreibirrelevanter Lesefähigkeiten in Abhängigkeit von den verwendeten schriftlichen Instruktionstexten zeigen, zumindest nicht, solange diese Instruktionstexte in einem gewissen für die Altersgruppe didaktisch angemessenen Schwierigkeitsspektrum anzusiedeln sind. Auch die Evidenzen zur Textmusterspezifität von Schreibkompetenzen sind relevant für die Unterrichtspraxis, so etwa im Hinblick darauf, dass für die Diagnose genereller Schreibfähigkeiten Aufgaben mehrerer Textmuster eingesetzt und entsprechende Kompetenzen abgeprüft werden sollten. Ebenso können die Resultate zur zweidimensionalen internen Struktur von Schreibkompetenzen nutzbar gemacht werden. Beispielsweise könnte alleine das Wissen darüber, dass es sich bei inhaltlich-stilistischen und orthografisch-grammatischen Schreibfähigkeiten um zwei differente, in weiten Teilen unabhängige Fähigkeiten handelt, die Wirksamkeit durch Aspekte der sprachlichen Richtigkeit bedingter Halo-Effekte bei der inhaltlichen und stilistischen Textbeurteilung reduzieren. Dies trifft vor allem dann zu, wenn man davon ausgeht, dass diese Halo-Effekte (zumindest zum Teil) auf expliziten und impliziten Annahmen über die kognitive Fähigkeitsstruktur der Schreibenden beruhen (Figueredo & Varnhagen, 2005; Kreiner et al., 2002; Varnhagen, 2000; vgl. Kapitel 8.2.2., 8.4.2. & 8.9.).

9.5. Ausblick

Wie bereits in Kapitel 9.2. erwähnt, konnten im Rahmen dieser Arbeit nur einige Aspekte der umfangreicheren Fragestellung „Wie valide ist das vorgestellte Schreibkompetenzmessungsverfahren?“ (bzw. entsprechende Formulierungen, je nach zugrunde gelegtem Validitätskonzept) untersucht werden. Darüber hinaus bestehen jedoch im Rahmen dieses Themenkomplexes weitere Unterfragestellungen, welche noch nicht im Rahmen dieser Arbeit oder anderer Arbeiten beantwortet wurden.

So wurde bereits in Kapitel 6.7. diskutiert, dass im Rahmen der Normierungsstudie aufgrund des Sets an verwendeten Aufgaben und des zugrunde gelegten Testdesigns keine Möglichkeit bestand, konkrete Aufgaben- oder Textsorteneffekte (unterhalb der Textmusterebene bzw.

auch quer dazu)[72] zu kontrollieren oder diese gezielt zu untersuchen. Dabei handelt es sich bei möglicher Aufgaben- und Textsortenmerkmalsensitivität jedoch um zwei Aspekte, welche auch für die Validität bei der Messung von Schreibkompetenzen relevant sind.

Im Rahmen dieser Arbeit konnte Evidenz dafür erbracht werden, dass Schreibkompetenzen textmusterspezifisch sind; ob jedoch auch auf anderen Ebenen der Textklassifikation Gemeinsamkeiten oder Unterschiede zwischen den erbrachten Schreibleistungen und den entsprechenden Konstrukten bestehen und auf welcher Klassifikationsebene solche Zusammenhänge oder Differenzen gefunden werden können, musste vorerst offen bleiben. Die Betrachtung dieser Aspekte könnte Aufschluss über mögliche weitere (textsorten-spezifische) Dimensionen des Konstrukts *Schreibkompetenz* liefern. Sollten sich die Schreib-leistungen in Abhängigkeiten von Textsorten oder Textsortenmerkmalen in relevantem Maße unterscheiden, müsste dies bei der Messung von *Schreibkompetenz* berücksichtigt werden, gemessene Schreibleistungswerte könnten nicht über die Textsorten der eingesetzten Aufgaben hinweg auf andere Textsorten oder auf ein textsortenunabhängiges Konstrukt *Schreibkompetenz* verallgemeinert werden. Astrid Neumann (2007), die, wie bereits in Kapitel 6.7. erläutert, das Schreiben von persönlichen und formellen Briefen untersuchte, fand Evidenzen für strukturelle Parallelitäten hinsichtlich dieser beiden Textsorten bzw. Textsortenmerkmale, konstatiert jedoch den Bedarf, „diese Ergebnisse für weitere Textsorten (…) [zu validieren], bevor sie über alle Textsorten (…) generalisiert werden“ (S. 202).

Zur Untersuchung möglicher Textsortensensitivität von Schreibkompetenzen bedarf es einer gezielten Aufgabenentwicklung, sodass ein Set an Aufgaben entsteht, welches das Textsorten-merkmal bzw. die Textsortenmerkmale unabhängig vom Textmuster variiert. Tabelle 9.5.1 illustriert dies exemplarisch an einem hypothetischen Aufgabenset mit den drei bekannten Textmustern, den beiden Textsorten *Brief* und *Zeitungsartikel* sowie dem Textsortenmerkmal *Formalität* als klassifikatorische Merkmale.[73] Darüber hinaus sollte für eine derartige Untersuchung im Rahmen des Testdesigns ein Verknüpfungsverfahren gewählt werden, welches Kontrastbedingungen, d. h. Aufgaben, welche sich nur in einem Merkmal unterscheiden, direkt miteinander verbindet; hinsichtlich des exemplarischen Aufgabensets aus Tabelle 9.5.1 wären beispielsweise für Aufgabe 1 direkte Verknüpfungen zu Aufgabe 2

[72] Vgl. Kapitel 6.2. zur Verwendung der Begriffe *Textmuster* und *Textsorte*.
[73] Die Kombination von *Zeitungsartikel* und *informell* mag auf den ersten Blick unüblich erscheinen, man denke jedoch beispielsweise an den Subtyp des Schülerzeitungsartikels, der je nach Situierung der Schülerzeitung (schüler- oder lehrergeleitet, nur Mitschüler oder auch Eltern und Lehrer als Leserschaft) sehr informell gehalten sein kann.

(unterschiedliches Testsortenmerkmal), 3 (unterschiedliche Textsorte), 5 und 9 (jeweils unterschiedliches Textmuster) in diesem Sinne ideal.

Tabelle 9.5.1: ***Beispiel für ein Aufgabenset zur Untersuchung verschiedener Textsortenmerkmale.***

Aufgabe	Textmuster	Textsorte	Textsortenmerkmal
1	argumentierend	Brief	förmlich
2	argumentierend	Brief	informell
3	argumentierend	Zeitungsartikel	förmlich
4	argumentierend	Zeitungsartikel	informell
5	informierend	Brief	förmlich
6	informierend	Brief	informell
7	informierend	Zeitungsartikel	förmlich
8	informierend	Zeitungsartikel	informell
9	narrativ	Brief	förmlich
10	narrativ	Brief	informell
11	narrativ	Zeitungsartikel	förmlich
12	narrativ	Zeitungsartikel	informell

Ebenso wie mögliche textsortenspezifische Effekte betreffen potentielle Aufgabeneffekte die Generalisierbarkeit von ermittelten Schreibkompetenzen. So fand Schoonen (2012) in einer niederländischen Schreibleistungsstudie, an welcher knapp 400 Schülerinnen und Schüler zu drei Messzeitpunkten (Jahrgangsstufe 8, 9 und 10) teilnahmen und je drei Schreibaufgaben bearbeiteten, dass etwa ein Drittel der Gesamtleistungsvarianz auf aufgabenspezifische Effekte (Faktor: Person × Aufgabe) entfiel. Schoonens weitere Analysen im Rahmen der Generalisierungstheorie (Cronbach, Gleser, Nanda & Rajaratnam, 1972) zeigten, dass selbst unter Einbeziehung mehrerer Aufgaben (bis zu vier) und mehrerer (bis zu drei) Raterurteile, ein kritischer Generalisierbarkeitswert von .8 (oder höher) nicht erreicht werden konnte (maximal erreichter Wert: .738; Schoonen, 2012, S. 14). Anzumerken ist jedoch, dass im Rahmen der Studie mehrere Textmuster und Textsorten (ohne systematische Variation) zum Einsatz kamen und diese Aspekte in der Analyse nicht berücksichtigt wurden, sodass sich reine (inhaltliche und/oder motivationale) Aufgabeneffekte nicht von textmuster- oder textsortenspezifischen Effekten trennen lassen.

Hinsichtlich möglicher thematischer Aufgabeneffekte bildet der Ansatz von Olinghouse et al. (2012), wie bereits in Kapitel 6 referiert, eine gute Grundlage für zukünftige Studien. Die Autoren konstruierten dabei drei verschiedene Aufgaben zu unterschiedlichen Textmustern, jedoch zum gleichen Großthema. Ziel der Autoren war es, unter Ausschaltung thematischer Variation bei dem Vergleich der aufgabenspezifischen Schreibleistungen Evidenzen für oder gegen die Textmusterspezifität von Schreibkompetenzen zu finden. Auch wenn das Ziel der Autoren, wie in Kapitel 6.1. erörtert, in dieser Weise nicht erreicht werden konnte, so stellt dieser Ansatz dennoch eine konzeptionelle Basis für die Untersuchung möglicher thematischer Aufgabeneffekte bereit. Für eine solche Untersuchung könnte in einem Teilset der Aufgaben das Textmuster bei gleichen Thema variiert werden, in einem dazu querliegenden Teilset das Thema unter Beibehalt des Textmusters (vgl. Tabelle 9.5.2). Auf diese Weise könnten Aufgaben- und Textmustereffekte sauber getrennt und unterschieden werden.

Tabelle 9.5.2: ***Beispiel für ein Aufgabenset zur Untersuchung von thematischen Aufgabeneinflüssen unter Kontrolle des Textmusters.***

Textmuster	**Thema 1** bspw. „Weltraum“ (Olinghouse et al. 2012)	**Thema 2**	**Thema 3**
argumentierend	Aufgabe A_1	Aufgabe A_2	Aufgabe A_3
informierend	Aufgabe I_1	Aufgabe I_2	Aufgabe I_3
narrativ	Aufgabe N_1	Aufgabe N_2	Aufgabe N_3

Eng mit der Frage nach thematischen Aufgabeneffekten verbunden ist die Frage hinsichtlich des motivationalen Einflusses bei der Messung von Schreibkompetenzen. Wie in Kapitel 2.2.1. erläutert, beinhaltet der in der heutigen Bildungsforschung vorherrschende Kompetenzbegriff auch motivationale Aspekte, *Kompetenz* wird verstanden als *Leistung* und *Bereitschaft* (Becker-Mrotzek & Schindler, 2007; Klieme & Hartig, 2008; Weinert, 2001). Dies steht im Einklang mit kognitiven Modellen der beim Schreiben involvierten Prozesse (Flower & Hayes, 1981; Hayes, 1996; Hayes & Flower, 1980; vgl. Kapitel 2.1.3.). Da mit einem solchen Kompetenzbegriff kein Anspruch auf maximale Leistung, sondern nur auf typische Leistung verbunden ist (Asseburg, 2011), scheinen motivationale Einflüsse auf den ersten Blick keine Gefährdung der Ergebnisse darzustellen, da die Bereitschaft zu schreiben Teil des Schreibkompetenzkonstruktes ist, sowohl im Test, als auch in externen Kontexten

(bspw. in der Schule oder in der alltäglichen Praxis). Dies setzt allerdings voraus, dass in der Testsituation dieselbe Motivationsstärke, eben die typische, vorliegt wie in testexternen Kontexten. Dass dies der Fall ist, muss jedoch erst geprüft werden. Dies kann beispielsweise durch den Einsatz bereits etablierter und erprobter Motivationsfragebögen, wie sie zur Validierung bereits in anderen Studien zur Kompetenzerfassung zum Einsatz gekommen sind, erfolgen, so etwa mittels des On-Line Motivation Questionnaire (OMQ), der auch in einer Kurzversion in der Haupttestung von PISA 2000 verwendet wurde (Boekaerts, 2002; Boekaerts & Otten, 1993; Crombach, Boekaerts & Voeten, 2003; Kunter et al., 2001), oder der QCM (Questionnaire on Current Motivation) bzw. FAM[74] (Fragebogen zur Erfassung aktueller Motivation in Lern- und Leistungssituationen) (Freund, Kuhl & Holling 2011; Rheinberg, Vollmeyer & Burns, 2001).

Während die Feststellung, ob die Motivationsstärke im Rahmen der Testsituation – vor allem im Rahmen von Low-stakes-Testungen, d. h. Testungen, die mit keinem persönlichen Gewinn, Verlust oder persönlichen Sanktionen für die Getesteten verbunden sind – hinreichend ähnlich zur testexternen Leistungsbereitschaft ist, eine Validierungsstrategie ist, die auch andere Kompetenzbereiche betrifft, scheint diese für den Bereich *Schreiben* jedoch von noch bedeutenderem Interesse. So konnten Sundre und Kitsantas (2004) für das Schreiben von Essays im Rahmen einer nichtrepräsentativen Studie (unter Beteiligung von 62 Psychologiestudierenden) verglichen mit der Bearbeitung eines Multiple-Choice-Tests einen gesteigerten Einfluss der Motivation auf die Schreibleistung nachweisen. Auch die am IQB im Rahmen der Normierungsstudie *Schreiben* erhobenen Daten sprechen aufgrund teilweise deutlich höherer Missing-Quoten im Vergleich zu Aufgaben anderer Kompetenzbereiche im Fach *Deutsch* für einen gesteigerten Einfluss motivationaler Aspekte auf die Bearbeitung freier Schreibaufgaben.

Ein zweiter gewichtiger Punkt, warum bei der Messung von Schreibkompetenzen motivationalen Aspekten eine besondere Bedeutung zukommt, ist, dass im praktischen Rahmen der Testungen, welche stets in einem zeitlich begrenzten Rahmen erfolgen müssen, lediglich der Einsatz einer überschaubaren Menge an Aufgaben pro Testperson möglich ist. Da diese Aufgaben sich nicht in mehrere Teilaufgaben und Items untergliedern, sondern die Aufgabenebene mit der Itemebene identisch ist, liegen pro Testperson lediglich wenige Item- bzw. Aufgabenwerte vor, anhand derer auf die Schreibkompetenz geschlossen wird. Nun fallen bei der Verwendung von Schreibleistungswerten nur weniger Aufgaben aufgaben-

[74] Bei FAM und QCM handelt es sich um inhaltlich (weitestmöglich) denselben Fragebogen, FAM ist lediglich die deutsche, QCM die englischsprachige Variante.

spezifische Effekte stärker ins Gewicht als im Rahmen einer Testung mit größerer Aufgaben- und Itemvarianz.

Dabei ist jede Aufgabe mit einer aufgaben- und individuumsspezifischen motivationalen Stimulanz (vgl. Kapitel 7.4.; Christmann & Groeben, 1999; Rosebrock, 2012) verbunden. Dies bedeutet auch, dass über die allgemeine Schreibleistungsmotivation hinaus die gewonnenen Schreibleistungswerte Einflüssen der aufgabenspezifischen Bearbeitungsmotivation unterliegen.

Die Erfassung aufgabenspezifischer motivationaler Aspekte kann ebenfalls durch den Einsatz geeigneter Fragebögen ermittelt werden, bspw. durch Fragen nach dem Interesse am Thema der Aufgabe, der generellen Relevanz und Wichtigkeit des Themas oder dessen Bezug zur Lebenswelt der Testperson. Ein entsprechender Fragenkatalog in Form eines etablierten Fragebogens steht bisher nicht zur Verfügung; zur Untersuchung der Fragestellung, in welchem Umfang Schreibleistungen durch die motivationale Stimulanz der Aufgabe beeinflusst sind, müsste ein solcher Fragebogen zunächst entwickelt und erprobt werden.

Diese drei hier hervorgehobenen Aspekte und Fragestellungen nach (i) dem motivationalen Einfluss sowie den möglichen Einflüssen von (ii) Aufgabenthematik und (iii) Textsorteneigenschaften, welche auf textmusterdifferenter Ebene zu verorten sind, können als Anstoß für zukünftige Untersuchungen im Bereich der Schreibkompetenzmessung betrachtet werden. Generell muss jedoch konstatiert werden, dass es sich auch bei diesen Punkten lediglich um einen Auszug handelt und nicht um eine vollständige Liste an offenen Fragen. Die Schreibkompetenzmessung im Large-Scale-Bereich ist noch relativ jung und nur wenige Studien wurden bisher in diesem Bereich durchgeführt; so ist zu wünschen, dass trotz Bedingungen des erhöhten personellen, zeitlichen und finanziellen Aufwands zahlreiche weitere Schreibleistungsstudien durchgeführt werden können, um die dargestellten sowie weitere offene Fragen zu untersuchen.

Literatur

Abraham, U. (1996). *StilGestalten. Geschichte und Systematik der Rede vom Stil in der Deutschdidaktik.* Tübingen: Niemeyer.

Abraham, U. (2014). Geschichte schulischen Schreibens. In H. Feilke & T. Pohl (Hrsg.), *Schriftlicher Sprachgebrauch – Texte verfassen* (S. 3–30). Baltmannsweiler: Schneider Hohengehren.

Abraham, W. (1995). *Deutsche Syntax im Sprachvergleich*. *Grundlegung einer typologischen Syntax des Deutschen.* Tübingen: Narr.

Adams, R. & Wu, M. (2010). *Modelling Polytomously Scored Items With The Rating Scale and Partial Credit Models*. Letzter Zugriff am 30.03.2015 unter http://www.acer.edu.au/files/Conquest-Tutorial-2-RatingScaleAndPartialCreditModels.pdf.

Adamzik, K. (1991). Forschungsstrategien im Bereich der Textsortenlinguistik. *Zeitschrift für Germanistik. Neue Folge 1*, 99–109.

Adamzik. K. (2004). *Textlinguistik. Eine einführende Darstellung.* Tübingen: Niemeyer.

Adamzik. K. (2008). Textsorten und ihre Beschreibung. In N. Janich (Hrsg.), *Textlinguistik. 15 Einführungen* (S. 145–176). Tübingen: Narr.

Aitchison, J. (1997). *Wörter im Kopf. Eine Einführung in das mentale Lexikon.* Tübingen: Niemeyer

Amstad, T. (1978). *Wie verständlich sind unsere Zeitungen?* Dissertation, Universität Zürich.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, *37*, 1–15.

Anderson, J. R. (1976). *Language, memory and thought.* Hillsdale, NJ: Erlbaum.

Androutsopoulos, J. (2007). Neue Medien – neue Schriftlichkeit. *Mitteilungen des deutschen Germanistenverbandes*, *1*(7), 72–97.

Antos, G. (1996). Textproduktion: Überlegungen zu einem fächerübergreifenden Schreib-Curriculum. In H. Feilke & P. R. Portmann (Hrsg.), *Schreiben im Umbruch. Schreibforschung und schulisches Schreiben* (S. 186–197). Stuttgart: Klett.

APA (2002): American Educational Research Association, American Psychological Association & National Council on Measurement in Education (2002). *Standards for educational and psychological testing.* Washington, D.C.: American Educational Research Association.

Arbuckle, J. L. & Wothke, W. (1999). *Amos 4.0 user's guide*. Marketing Department, SPSS Incorporated.

Arlt, F. & Beelitz, A. (1970). *Führungskräfte der Wirtschaft äußern sich zu Lehr-und Lernzielen der Hauptschule: Ergebnisse und Kommentierung einer Befragung*. Hannover: Schroedel.

Artelt, C., McElvany, N., Christmann, U., Richter, T., Groeben, N., Köster, J., Schneider, W., Stanat, P., Ostermeier, C., Schiefele, U., Valtin, R. & Ring, K. (2007). *Förderung von Lesekompetenz: Expertise.* Bonn, Berlin: Bundesministerium für Bildung und Forschung.

Asseburg, R. (2011). *Motivation zur Testbearbeitung in adaptiven und nicht-adaptiven Leistungstests.* Dissertation, Christian-Albrechts-Universität zu Kiel.

Augst, G., Disselhoff, K., Henrich, A., Pohl, T. & Völzing, P.-L. (2007). *Text-Sorten-Kompetenz. Eine echte Longitudinalstudie zur Entwicklung der Textkompetenz im Grundschulalter.* Frankfurt am Main et al.: Lang.

Austin, J. L. (1972). *Zur Theorie der Sprechakte (How to do things with Words).* Stuttgart: Reclam.

Bachman, L. & Palmer, A. (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

Bachmann, T. (2002). *Kohäsion und Kohärenz: Indikatoren der Schreibentwicklung. Zum Aufbau kohärenzstiftender Strukturen in instruktiven Texten von Kindern und Jugendlichen.* Innsbruck, Wien, München, Bozen: Studien Verlag.

Baddeley, A. D. (1997). *Human memory: Theory and practice*. Hove: Psychology Press.

Baddeley, A. D., Thomson, N. & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of verbal learning and verbal behavior*, *14*(6), 575–589.

Bamberger, R. & Vanecek, E. (1984). *Lesen - Verstehen - Lernen - Schreiben. Die Schwierigkeitsstufen von Texten in deutscher Sprache.* Wien: Jugend und Volk.

Baretta, L., Tomitch, L. M. B., MacNair, N., Lim, V. K. & Waldie, K. E. (2009). Inference making while reading narrative and expository texts: an ERP study. *Psychology & Neuroscience*, *2*(2), 137–145.

Bauer, B. A. (1981). *A Study of the Reliabilities and the Cost-Efficiencies of Three Methods of Assessment for Writing Ability*. Urbana, IL: University of Illinois.

Baumann, K.-D. (1992). *Integrative Fachtextlinguistik*. Tübingen: Narr.

Baumert, J., Bos, W. & Lehmann, R. (Hrsg.) (2000). *Dritte Internationale Mathematik und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn.* Opladen: Leske + Budrich.

Baumert, J., Stanat, P. & Demmrich, A. (2001). PISA 2000: Untersuchungsgegenstand, theoretische Grundlagen und Durchführung der Studie. In Deutsches Pisa-Konsortium (Hrsg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 15–68). Opladen: Leske + Budrich.

Baumert, J., Stanat, P. & Watermann, R. (2006). Schulstruktur und die Entstehung differenzieller Lern-und Entwicklungsmilieus. In J. Baumert, P. Stanat & R. Watermann (Hrsg.), *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit. Vertiefende Analysen im Rahmen von PISA 2000* (S. 95–188). Wiesbaden: VS Verlag für Sozialwissenschaften.

Baurmann, J. & Weingarten, R. (Hrsg.) (1995). *Schreiben: Prozesse, Prozeduren und Produkte*. Opladen: Westdeutscher Verlag.

Bayer, K. (2007). *Argument und Argumentation: logische Grundlagen der Argumentationsanalyse*. Göttingen: Vandenhoeck & Ruprecht.

Bechger, T. M., Maris, G. & Hsiao, Y. P. (2007). Assessing the size of halo-effects in performance based tests and a practical solution to avoid halo-effects. *Measurement and Research Department Reports*, *2007-2*. Arnheim: CITO, National Institute for Educational Measurement.

Beck, O. (1974). *Kriterien zur Aufsatzbeurteilung*. V. Hase und Koehler: Mainz

Beck, O. (1979). *Theorie und Praxis der Aufsatzbeurteilung: Forschungsstand, Wege der Objektivierung und Leistungsförderung: Ein Handbuch für Lehrende und Studierende*. Bochum: Ferdinand Kamp.

Becker-Mrotzek, M. & Böttcher, I. (2014). *Schreibkompetenz entwickeln und beurteilen. [Sekundarstufe I/II]* (5. Auflage). Berlin: Cornelsen.

Becker-Mrotzek, M., Jost, J., Knopp, M. & Grabowski, J. (2011). *Teilkomponenten von Schreibkompetenz: Diagnose und Förderung.* Präsentation GFD Fachtagung: Formate fachdidaktischer Forschung, Berlin.

Becker-Mrotzek, M. & Schindler, K. (2007). Schreibkompetenz modellieren. In M. Becker-Mrotzek & K. Schindler (Hrsg.), *Texte schreiben*. *Kölner Beiträge zur Sprachdidaktik, 5* (S. 7–26). Köln: Gilles & Francke.

Behrens, U. & Krelle, M. (2011). Schülertexte beurteilen im Licht von Bildungsstandards, Kompetenzrastern und Unterrichtsalltag. *éro 93*, 167–183.

Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Loudermilk, B. L., Kromrey, J. D. & Ferron, J. M. (2010). *Dancing the sample size limbo with mixed models: How low can you go.* Proceedings of the SAS Global Forum, April 2010. Cary, NC: SAS Institute Inc.

Bereiter, C. (1980). Development in Writing. In L. W. Gregg & E. R. Steinberg, (Eds.), *Cognitive Processes in Writing* (pp. 73–93). Hillsdale: Erlbaum.

Berkowitz, S. & Taylor, B. (1981). The effects of text type and familiarity on the nature of information recalled by readers. *Directions in reading: Research and instruction*, 157–161.

Berman, R. A. & Nir-Sagiv, B. (2007). Comparing narrative and expository text construction across adolescence: A developmental paradox. *Discourse processes*, *43*(2), 79–120.

Berninger, V. W., Abbott, R. D., Jones, J., Wolf, B. J., Gould, L., Anderson-Youngstrom, M., Shimada, S. & Apel, K. (2006). Early development of language by hand: Composing, reading, listening, and speaking connections; three letter-writing modes; and fast mapping in spelling. *Developmental Neuropsychology*, *29*(1), 61–92.

Best, K.-H. (2005). Wortlänge. In R. Köhler, G. Altmann & R. G. Piotrowski (Hrsg.), *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch* (S. 260–273). Berlin, New York: de Gruyter.

Best, K.-H. (2006a). Sind Wort- und Satzlänge brauchbare Kriterien der Lesbarkeit von Texten? In S. Wichter & A. Busch (Hrsg.), *Wissenstransfer – Erfolgskontrolle und Rückmeldungen aus der Praxis* (S. 21–31). Frankfurt am Main et al.: Lang.

Best, K.-H. (2006b). Wortlängen im Deutschen. *Göttinger Beiträge zur Sprachwissenschaft 13*, 23–49.

BIFIE (Hrsg.) (2012). *Themenheft für den Kompetenzbereich „Verfassen von Texten". Deutsch, Lesen, Schreiben. Volksschule Grundstufe I + II.* Graz: Leykam.

Birkel, P. (2003). Aufsatzbeurteilung – ein altes Problem neu untersucht. *Didaktik Deutsch*, *9*(15), 46–63.

Birkel, P. & Birkel, C. (2002). Wie einig sind sich Lehrer bei der Aufsatzbeurteilung? Eine Replikationsstudie zur Untersuchung von Rudolf Weiss. *Psychologie in Erziehung und Unterricht*, *49*, 219–224.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: MIT Press.

Björnsson, C. H. (1968). *Läsbarhet*. Stockholm: Liber.

Bley-Vroman, R. (1989). What is the logical problem of foreign language learning? In: S. M. Gass & J. Schachter (Eds.), *Linguistic Perspectives on Second Language Acquisition* (S. 41–68). Cambridge: Cambridge University Press.

Blömeke, S. (2013). *Validierung als Aufgabe im Forschungsprogramm „Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor"* (KoKoHs Working Papers, Nr. 2). Berlin & Mainz: Humboldt-Universität & Johannes Gutenberg-Universität.

Boekaerts, M. (2002). The on-line motivation questionnaire: A self-report instrument to assess students' context sensitivity. In P. R. Pintrich & M. L. Maehr (Eds.), *Advances in Motivation and Achievement, Vol. 12, New Directions in Measures and Methods* (pp. 77–120). Oxford, UK: Elsevier.

Boekaerts, M. & Otten, R. (1993). Handlungskontrolle und Lernanstrengung im Schulunterricht. *Zeitschrift für Pädagogische Psychologie, 7*(2/3), 109–116.

Boettcher, W., Firges, J., Sitta, H., Tymister, H. J. (1973). *Schulaufsätze - Texte* für *Leser*. Düsseldorf: Schwann.

Böhme, K. (2012). *Methodische und didaktische Überlegungen sowie empirische Befunde zur Erfassung sprachlicher Kompetenzen im Deutschen.* Dissertation, Humboldt-Universität zu Berlin.

Böhme, K., Bremerich-Vos, A. & Robitzsch, A. (2009). Aspekte der Kodierung von Schreibaufgaben: Vergleich holistischer und analytischer Kodierungen unter besonderer Berücksichtigung der Interraterreliabilität. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 290–329). Weinheim: Beltz.

Böhme, K., Richter, D., Stanat, P., Pant, H. A. & Köller, O. (2012). Kapitel 1: Die länderübergreifenden Bildungsstandards in Deutschland. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik, Ergebnisse des IQB-Ländervergleichs* (S. 11–18). Münster, New York, München, Berlin: Waxmann.

Böhme, K., Tiffin-Richards, S. P., Schipolowski, S. & Leucht, M. (2010). Migrationsbedingte Disparitäten bei sprachlichen Kompetenzen. In O. Köller, M. Knigge & B. Tesch (Hrsg.), *Sprachliche Kompetenzen im Ländervergleich* (S. 203–225). Münster: Waxmann.

Borsboom, D., Cramer, A., Kievit, R., Zand Scholten, A. & Franic, S. (2009). The end of construct validity. In R. Lissitz (Ed.), *The concept of validity* (135–170). Charlotte, NC: Information Age Publishers.

Borsboom, D. & Markus, K. A. (2013). Truth and Evidence in Validity Theory. *Journal of Educational Measurement, 50*, 110–114.

Borsboom, D., Mellenbergh, G. J. & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071.

Borsche, T. (1989). Die innere Form der Sprache. Betrachtungen zu einem Mythos der Humboldt-Herme(neu)tik. In H. W. Scharf (Hrsg.), *Humboldts Sprachdenken* (S. 47–65). Essen: Hobing.

Bortz, J. (2005). *Statistik für Human-und Sozialwissenschaftler* (6. Auflage). Heidelberg: Springer.

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation: für Human- und Sozialwissenschaftler* (2. Auflage). Berlin: Springer.

Brandstätter, E. (1999). Konfidenzintervalle als Alternative zu Signifikanztests. *Methods of Psychological Research Online*, *4*(2), 1–17.

Bransford, J. D., Barclay, J. R. & Franks, J. J. (1972). Sentence memory: A constructive versus interpretive approach. *Cognitive psychology*, *3*(2), 193–209.

Bransford, J. D. & Franks, J. J. (1971). The abstraction of linguistic ideas. *Cognitive psychology*, *2*(4), 331–350.

Braunschweig Verlag (2014). *Zeichensetzung: Die wichtigsten Unterschiede zwischen der englischen und deutschen Interpunktion.* Letzter Letzter Zugriff am 30.03.2015 unter http://www.braunschweig-verlag.de/zeichensetzung-deutsch-englisch-vergleich.html.

Bremerich-Vos, A., Behrens, U., Böhme, K., Krelle, M., Neumann, D., Robitzsch, A., Schipolowski, A. & Köller, O. (2010). Kompetenzstufenmodelle für das Fach Deutsch. In O. Köller, M. Knigge & B. Tesch (Hrsg.), *Sprachliche Kompetenzen im Ländervergleich. Überprüfung der Bildungsstandards in den Fächern Deutsch und erste Fremdsprache in der neunten Jahrgangsstufe* (S. 37–50). Münster: Waxmann.

Bremerich-Vos, A., Böhme, K., Krelle, M., Weirich, S., Köller, O. (2012). Kompetenzstufenmodelle im Fach Deutsch. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik* (S. 56–71). Münster: Waxmann.

Bremerich-Vos, A., Böhme, K. & Robitzsch, A. (2009). Sprachliche Kompetenzen im Fach Deutsch. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (198–218). Weinheim: Beltz.

Briggs, D. (1970). The influence of handwriting on assessment. *Educational Research 13*, 50–55.

Brinker, K. (1985). *Linguistische Textanalyse: Eine Einführung in Grundbegriffe und Methoden (= Grundlagen der Germanistik; 29)*. Berlin: Schmidt.

Bryant, F. B. (2000). Assessing the Validity of Measurement. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding MORE multivariate statistics.* Washington D.C.: American Psychological Association.

Bucher, H. J. (1999). Medien-Nachbarwissenschaften III: Linguistik. In: J. F. Leonhard (Hrsg), *Medienwissenschaft. Ein Handbuch zur Entwicklung der Medien und Kommunikationsformen*, *1* (S. 287–309). Berlin, New York: de Gruyter.

Buchhaas-Birkholz, D. (2009). Die „empirische Wende" in der Bildungspolitik und in der Bildungsforschung. Zum Paradigmenwechsel des BMBF im Bereich der Forschungsförderung. *Erziehungswissenschaft*, *20*(39), 27–33.

Budde, M. (2012). Über Sprache reflektieren. *Unterricht in sprachheterogenen Lerngruppen-Fernstudieneinheit*. Kassel: Kassel University Press.

Bühler, K. (1934). *Sprachtheorie. Die Darstellungsfunktion der Sprache.* Jena: Fischer.

Bühner, M. (2004). *Einführung in die Test-und Fragebogenkonstruktion.* München: Pearson Studium.

Bundesministerium für Bildung und Forschung (BMBF) (2012). *Lesen & Schreiben – Mein Schlüssel zur Welt" Leitfaden zur Umsetzung von Aktionen im Rahmen der Kampagne „Lesen & Schreiben – Mein Schlüssel zur Welt" sowie Tipps und Hinweise für die begleitende Presse- und Medienarbeit.* Berlin. Letzter Zugriff am 30.03.2015 unter http://www.mein-schlüssel-zur-welt.de/_files/BMBF_Aktionsleitfaden.pdf.

Byrne, B. M. (2001). Structural equation modeling with AMOS, EQS, and LISREL: Comparative approaches to testing for the factorial validity of a measuring instrument. *International Journal of Testing*, *1*(1), 55–86.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.

Caravolas, M. (2004). Spelling development in alphabetic writing systems: A cross-linguistic perspective. *European Psychologist*, *9*(1), 3–14.

Caravolas, M. & Landerl, K. (2010). The influences of syllable structure and reading ability on the development of phoneme awareness: A longitudinal, cross-linguistic study. *Scientific Studies of Reading*, *14*(5), 464–484.

Carlman, N. (1985). *Variations in the writing performance of grade 12 students: Differences by mode and topic.* ERIC Document Reproduction Services No. ED269766.

Carroll, S. E. (2001). *Input and Evidence. The Raw Material of Second Language Acquisition*. Amsterdam, Philadelphia: Benjamins.

Caspari, D., Kleppin, K. & Grotjahn, R. (2010). Testaufgaben und Lernaufgaben. In R. Porsch (Hrsg.), *Standardbasierte Testentwicklung und Leistungsmessung Französisch in der Sekundarstufe I* (S. 46–68). Münster: Waxmann.

Chalifour, C. L. & Powers, D. E. (1989). The relationship of content characteristics of GRE analytical reasoning items to their difficulties and discriminations. *Journal of Educational Measurement*, *26*(2), 120–132.

Chase, C. (1968). The impact of some obvious variables on essay test scores. *Journal of Educational Measurement*, *5*, 315–318.

Chita, A. (2008). *Bewertungskriterien schriftlicher Lernerproduktionen B2 und C1 und ihre Validität.* Dissertation, Universität Augsburg.

Chomsky, N. (1962). Explanatory Models in Linguistics. In: E. Nagel, P. Suppes & A. Traski (Eds.), *Logic, Methodology and Philosophy of Science* (pp. 528–555). Stanford, CA: Stanford University Press.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, Massachusetts: MIT Press.

Chomsky, N. (2000). *The architecture of language*. New Delhi: Oxford University Press.

Christmann, U. & Groeben, N. (1999). Psychologie des Lesens. In B. Franzmann, K. Hasemann, D. Löffler & E. Schön (Hrsg.), *Handbuch Lesen* (S. 145–223). München: Saur.

Cicchetti, D. V. & Prusoff, B. A. (1983). Reliability of depression and associated clinical symptoms. *Archives of General Psychiatry*, *40*(9), 987–990.

Cicchetti, D. V. & Sparrow, S. S. (1981). Development of criteria for establishing the interrater reliability of specific items in a given inventory: applications to assessment of adaptive behavior. *American Journal of Mental Deficiency, 86*, 127–137.

Cizek, G. J. & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. London: SAGE Publications Ltd.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin 114*, 494–509.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cohen, J. & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cohen, R. A., Sparling-Cohen, Y. A. & O'Donnell, B. F. (1993). *The neuropsychology of attention*. New York: Plenum Press.

Coltheart, M. (1978). Lexical Access in Simple Reading Tasks. In G. Underwood (Ed.), *Strategies of Information Processing* (pp. 151–216). London: Academic Press.

Conrad, R. (Hrsg.) (1985). Lexikon sprachwissenschaftlicher Termini. Leipzig: VEB Bibliographisches Institut.

Crombach, M. J., Boekaerts, M. & Voeten, M. J. (2003). Online measurement of appraisals of students faced with curricular tasks. *Educational and psychological measurement*, *63*(1), 96–111.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (3–17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (147–171). Urbana, IL: University of Illinois Press.

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Crotti, C. & Osterwalder, F. (Hrsg.) (2008). *Das Jahrhundert der Schulreformen. Internationale und nationale Perspektiven, 1900–1950*. Bern, Stuttgart, Wien: Haupt.

Cummins, J. (2006). Sprachliche Interaktion im Klassenzimmer. Von zwangsweise auferlegten zu kooperativen Formen von Machtbeziehungen. In P. Mecheril & T. Quehl, (Hrsg.), *Die Macht der Sprachen* (S. 36–62). Münster, New York: Waxmann.

Cutler, A. & Clifton, C. (1999). Comprehending spoken language: A blueprint of the listener. In C. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 123–166). Oxford: Oxford University Press.

Daneš, F. (1970). Zur linguistischen Analyse der Textstruktur. *Folia linguistica*, *4*(1–2), 72–78.

Danks, J. H. & End, L. J. (1987). Processing strategies for reading and listening. In R. Horowitz & S. J. Samuels (Eds.), *Comprehending oral and written language* (pp. 271–294). San Diego, London: Academic Press.

de Beaugrande, R.-A. & Dressler, W. U. (1981). *Einführung in die Textlinguistik*. Tübingen: Niemeyer.

Deinzer, R. (2007). *Allgemeine Grundlagen wissenschaftlichen Arbeitens in der Medizin: ein Leitfaden für die empirische Promotion und Habilitation*. Stuttgart: Kohlhammer.

Devitt, A. J. (2008). *Writing genres*. Carbondale, IL: Southern Illinois University Press.

Diakidoy, I. A. N., Stylianou, P., Karefillidou, C. & Papageorgiou, P. (2005). The relationship between listening and reading comprehension of different types of text at increasing grade levels. *Reading Psychology, 26*(1), 55–80.

Draxler, D. (2005). *Aufgabendesign und basismodellorientierter Physikunterricht.* Dissertation, Universität Duisburg-Essen.

Driemeyer, W., Spehr, A., Yoon, D., Richter-Appelt, H. & Briken, P. (2013). Comparing sexuality, aggressiveness, and antisocial behavior of alleged juvenile sexual and violent offenders. *Journal of forensic sciences 58*(3), 711–718.

DuBay, W. H. (2007). *Smart Language: Readers, Readability, and the Grading of Text.* Letzter Zugriff am 30.03.2015 unter http://www.nald.ca/library/learning/smartlang/smartlang.pdf.

DuBay, W. H. (2004). *The principles of readability.* California: Impact Information.

Ehlich, K. (Hrsg.) (1980). *Erzählen im Alltag*. Frankfurt am Main: Suhrkamp.

Ehmke, T., Klieme, E. & Stanat, P. (2013). Veränderungen der Lesekompetenz von PISA 2000 nach PISA 2009. Die Rolle von Unterschieden in den Bildungswegen und in der Zusammensetzung der Schülerschaft. In N. Jude & E. Klieme (Hrsg.), *PISA 2009-Impulse für die Schul-und Unterrichtsforschung (Zeitschrift für Pädagogik, Beiheft; 59)* (S. 132–150). Weinheim et al.: Beltz.

Elke, A. (2007). *Unterrichten zur Förderung von selbst reguliertem Lernen in der Berufsbildung. Lehrervoraussetzung, Lehrerentwicklung und Perspektiven – Eine Interventionsstudie.* Dissertation, Universität Basel.

Engelen, B. (1974). Vorbemerkung zu einem an Kommunikationssituationen orientierten 'Aufsatz'unterricht. In A. Schau (Hrsg.), *Von der Aufsatzkritik zur Textproduktion. Beiträge zur Neugestaltung schriftlicher Sprachproduktion.* (S. 240–247). Baltmannsweiler: Schneider.

Engelhard, G., Gordon, B. & Gabrielson, S. (1992). The influences of mode of discourse, experiential demand, and gender on the quality of student writing. *Research in the Teaching of English, 26*, 315–335.

Engelhard, G., Gordon, B., Gabrielson, S. & Walker, E. V. S. (1994). Writing tasks and gender: Influences on writing quality of black and white students. *Journal of Educational Research, 87*, 197–209.

Engelkamp, J. (1995). Mentales Lexikon: Struktur und Zugriff. In G. Harras (Hrsg.), *Die Ordnung der Wörter. Kognitive und lexikalische Strukturen* (S. 99–119). Berlin: de Gruyter.

Ernst, F. (2011a). Lesbarkeit von Rechnungswesenbüchern an kaufmännischen Berufsschulen. *Zeitschrift für Berufs-und Wirtschaftspädagogik*, *107*(3), 408–423.

Ernst, F. (2011b). Lesbarkeit von Schulbüchern: Wie die Lesbarkeit den Lernerfolg beeinflusst und wie man Lesbarkeitsformeln anwendet. *Wirtschaft und Erziehung* (1–2), 17–20.

Erpenbeck, J. & Rosenstiel, L. v. (2007). Einführung. In J. Erpenbeck & L. Rosenstiel (Hrsg.), *Handbuch Kompetenzmessung. Erkennen, verstehen und bewerten von Kompetenzen in der betrieblichen, pädagogischen und psychologischen Praxis* (S. XVII–XLVI). Stuttgart: Schaffer-Poeschel.

Ervin-Tripp, S. M. (1974). Is second language learning like the first. *Tesol Quarterly*, 111–127.

ETS (2010). *User Guide Speaking and writing.* Princeton, NJ. Letzter Zugriff am 30.03.2015 unter http://www.ets.org/s/toeic/pdf/toeic_sw_score_user_guide.pdf.

Falke, Lisa. (2008). *Measures* of *reading comprehension*: *The effects of text type and time limits on students' performance.* Master Thesis, University of North Texas.

Feilke, H. (2006). Entwicklung schriftlich-konzeptueller Fähigkeiten. In U. Bredel, H. Günther, P. Klotz, G. Siebert-Ott & J. Ossner (Hrsg.), *Didaktik der deutschen Sprache. Ein Handbuch* (1. Teilband, 2. Auflage, S. 178–192). Paderborn et al.: Schöningh.

Feilke, H. (2014). Schriftliches Berichten. In H. Feilke & T. Pohl (Hrsg.), *Schriftlicher Sprachgebrauch – Texte verfassen* (S. 233–251). Baltmannsweiler: Schneider Hohengehren.

Ferguson, C. A. (1993). The language factor in national development. *Anthropological linguistics 4*(1), 23–27.

Fienemann, J. (2006). *Erzählen in zwei Sprachen*. Münster: Waxmann.

Figueredo, L. & Varnhagen, C. K. (2005). Didn't you run the spell checker? Effects of type of spelling error and use of a spell checker on perceptions of the author. *Reading Psychology*, *26*(4–5), 441–458.

Fischer, C. (2010). *Texte, Gattungen, Textsorten und ihre Verwendung in Lesebüchern.* Dissertation, Justus-Liebig-Universität Gießen.

Fisher, R. A. (1921). On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. *Metron*, *1*, 3–32.

Fishman, J., Lunsford, A., McGregor, B. & Otuteye, M. (2005). Performing writing, performing literacy. *College composition and communication 57*(2), 224–252.

Fisseni, H. J. (2004). *Lehrbuch der psychologischen Diagnostik: mit Hinweisen zur Intervention* (3. Auflage). Göttingen: Hogrefe.

Fitzgerald, J. & Shanahan, T. (2000). Reading and writing relations and their development. *Educational Psychologist, 35*(1), 39–50.

Fix, M. (2006). *Texte schreiben. Schreibprozesse im Deutschunterricht.* Paderborn, München, Wien, Zürich: Schöningh.

Fix, M. & Melenk, H. (2002). *Schreiben zu Texten – Schreiben zu Bildimpulsen*. Baltmannsweiler: Schneider Hohengehren.

Fix, U. (2008a). Text und Textlinguistik. In N. Janich (Hrsg.), *Textlinguistik. 15 Einführungen* (S. 15–34). Tübingen: Narr.

Fix, U. (2008b). *Texte und Textsorten: sprachliche, kommunikative und kulturelle Phänomene.* Berlin: Frank & Timme.

Fleiss, J. L. & Cohen, J. (1973). The equivalence of weighted Kappa and the Intraclass Correlation Coefficient as measures of reliability. *Educational and Psychological Measurement, 33*, 613–619.

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, *32*(3), 221.

Flower, L. & Hayes, J. R. (1981). A cognitive process theory of writing. *College composition and communication 32*, 365–387.

Flynn, S. (1996). A parameter-setting approach to second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *Handbook of Second Language Acquisition* (pp. 124–158). San Diego, CA: Academic Press.

Frazier, L. (1978). *On comprehending sentences: Syntactic parsing strategies.* Doctoral dissertation, University of Connecticut.

Frazier, L. & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive psychology*, *14*(2), 178–210.

Frederiksen, C. H. (1975). Representing logical and semantic structure of knowledge acquired from discourse. *Cognitive psychology*, *7*(3), 371–458.

Freedle, R. & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing*, *10*(2), 133–170.

Freund, P. A., Kuhn, J.-T., Holling, H. (2011). Measuring current achievement motivation with the QCM: Short form development and investigation of measurement invariance. *Personality and Individual Differences*, *51*(5), 629–634.

Friedrich, B. (1988). Informationen zur Durchsicht und Korrektur der Lehrpläne für den Muttersprachunterricht der Klassen 5 – 10. *Deutschunterricht 41,* Heft 4, 149–153.

Frith, U., Wimmer, H. & Landerl, K. (1998). Differences in phonological recoding in German- and English-speaking children. *Scientific Studies of Reading, 2*, 31–54.

Fuchs, H. W. (2009). Neue Steuerung – neue Schulkultur? *Zeitschrift für Pädagogik*, *55*(3), 369–380.

Gansel, C. & Jürgens, F. (2008). Textgrammatische Ansätze. In N. Janich (Hrsg.), *Textlinguistik. 15 Einführungen* (S. 55–83). Tübingen: Narr.

Gansel, C. & Jürgens, F. (2009). *Textlinguistik und Textgrammatik.* Göttingen: Vandenhoeck & Ruprecht.

Gantefort, C. (2013). *Schriftliches Erzählen mehrsprachiger Kinder: Entwicklung und sprachenübergreifende Fähigkeiten*. Münster: Waxmann.

Gardner, M. J. & Altman, D. G. (1986). Confidence intervals rather than P values: estimation rather than hypothesis testing. *British Medical Journal*, *292*(6522), 746–750.

Gärtner, H. & Pant, H. A. (2011). Validierungsstrategien für Verfahren und Ergebnisse von Schulinspektion. In S. Müller, M. Pietsch & W. Bos (Hrsg.), *Schulinspektion in Deutschland. Eine Zwischenbilanz in empirischer Sicht* (S. 9–32). Münster: Waxmann.

Gätje, O. (2013). Schreiben in der Sekundarstufe I. In S. Gailberger & F. Wietzke (Hrsg.), *Handbuch kompetenzorientierter Deutschunterricht* (S. 232–254). Weinheim, Basel: Beltz.

Gernsbacher, M. A., Varner, K. R. & Faust, M. E. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(3), 430.

Ginther, A. & Stevens, J. (1998). Language background, ethnicity, and the internal construct validity of the advanced placement Spanish language examination. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 169–194). Mahawah, NJ: Erlbaum.

Gipper, H. (Hrsg.) (1959). *Sprache – Schlüssel zur Welt: Festschrift für Leo Weisgerber*. Düsseldorf: Pädagogischer Verlag Schwann.

Grabowski, J., Blabusch, C. & Lorenz, T. (2007). Welche Schreibkompetenz? – Handschrift und Tastatur in der Hauptschule. In M. Becker-Mrotzek & K. Schindler (Hrsg.), *Texte schreiben. Kölner Beiträge zur Sprachdidaktik, 5* (S. 41–61). Köln: Gilles & Francke.

Groeben, N. (1982). *Leserpsychologie: Textverständnis – Textverständlichkeit.* Münster: Aschendorff.

Grohnfeldt, M. (Hrsg.). (2007). *Lexikon der Sprachtherapie*. Stuttgart: Kohlhammer.

Grotlüschen, A. & Riekmann, W. (2012). *Funktionaler Analphabetismus in Deutschland. Ergebnisse der ersten leo. Level-One Studie.* Münster, New York, München, Berlin: Waxmann.

Grzesik, J. & Fischer, M. (1984). *Was leisten Kriterien für die Aufsatzbeurteilung? Theoretische, empirische und praktische Aspekte des Gebrauchs von Kriterien und der Mehrfachbeurteilung nach globalem Ersteindruck*. Opladen: Westdeutscher Verlag.

Gülich, E. & Hausendorf, H. (2000). Vertextungsmuster Narration. In K. Brinker, G. Antos, W. Heinemann & S. Sager (Hrsg.), *Text-und Gesprächslinguistik/Linguistics of Text and Conversation* (1. Halbband, S. 369–385). Berlin: de Gruyter.

Hadenfeldt, J. C. & Neumann, K. (2012). Die Erfassung des Verständnisses von Materie durch Ordered Multiple Choice Aufgaben. *Zeitschrift für Didaktik der Naturwissenschaften, 18,* 317–338.

Hambleton, R. K. & Pitoniak, M. J. (2006). Setting performance standards. In R. Brennan (Ed.), *Educational measurement, 4th ed.* (pp. 433–470). Washington D.C.: American Council on Education and Westport, CT: Praeger Publishers.

Hamp-Lyons, L. (1991). Pre-text: Task-related influences on the writer. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 87–107). Norwood, NJ: Ablex.

Hamp-Lyons, L. (1996). The challenge of second-language writing assessment. In E. M. White, W. D. Lutz & S. Kamusikiri (Eds.), *Assessment of writing: politics, policies, practices* (pp. 226–240). New York: Modern Language Association of America.

Harris, D. P. (1969). *Testing English as a second language*. New York: McGraw-Hill.

Harsch, C. (2005). *Der Gemeinsame europäische Referenzrahmen für Sprachen: Leistung und Grenzen. Die Bedeutung des Referenzrahmens im Kontext der Beurteilung von Sprachvermögen am Beispiel des semikreativen Schreibens im DESI-Projekt.* Dissertation, Universität Augsburg.

Harsch, C., Neumann, A., Lehmann, R. & Schröder, K. (2007). Schreibfähigkeit. In E. Klieme B. & Beck (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)* (S. 42–66). Weinheim, Basel: Beltz.

Hartig, J. (2004). Methoden zur Bildung von Kompetenzstufenmodellen. In H. Moosbrugger, W. Rauch & D. Frank (Hrsg.), *Qualitätssicherung im Bildungswesen.* Frankfurt am Main: Arbeiten aus dem Institut der Johann Wolfgang Goethe-Universität, Heft 2004/03.

Hartig, J., Frey, A. & Jude, N. (2008). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 135–163). Heidelberg: Springer.

Hartig, J., Jude, N. & Wagner, W. (2008). Methodische Grundlagen der Messung und Erklärung sprachlicher Kompetenzen. In DESI-Konsortium (Hrsg.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (S. 34–54). Weinheim u. a.: Beltz.

Hartig, J. & Kühnbach, O. (2006). Schätzung von Veränderung mit „plausible values " in mehrdimensionalen Rasch-Modellen. In A. Ittel & H. Merkens (Hrsg), *Veränderungsmessung und Längsschnittstudien in der empirischen Erziehungswissenschaft.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Hartmann, P. (1971). Texte als linguistisches Objekt. In W.-D. Stempel (Hrsg.), *Beiträge zur Textlinguistik* (S. 9–29). München: Fink.

Hartmann, W. & Jonas, H. (Hrsg.) (1996). *Deutschunterricht im Umbruch. Die Aufsatzstudie Ost von 1991.* Frankfurt am Main: Lang.

Hartmann, W. & Lehmann, R. H. (1987). *The Hamburg Study of Achievement in Written Composition: National Report for the IEA International Study of Achievement in Written Communication, Part I: Method and Findings.* Hamburg: University of Hamburg.

Harweg, R. (1968). *Pronomina und Textkonstitution.* München: Fink.

Haueis, E. (2006). Formen schriftlicher Texte. In U. Bredel, H. Günther, P. Klotz, J. Ossner & G. Siebert-Ott (Hrsg.), *Didaktik der deutschen Sprache. Ein Handbuch.* (1. Teilband, 2. Auflage, S. 224–236). Paderborn: Schöningh.

Hayes, J. R. (1996). A framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Hrsg.), *The science of writing. Theories, Methods, Individual Differences and Application* (pp. 1–27). Mahwah, NJ: Erlbaum.

Hayes, J. R. & Flower, L. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp 3–30). Hillsdale: Erlbaum.

Heim, S., Keil, A. & Ihssen, N. (2006). Der Zusammenhang zwischen zeitlicher Aufmerksamkeitsallokation und Lese-Rechtschreibleistungen im frühen Sekundarschulalter. *Zeitschrift für Psychologie/Journal of Psychology*, *214*(4), 196–206.

Heinemann, M. & Heinemann, W. (2002). *Grundlagen der Textlinguistik. Interaktion – Text – Diskurs.* Tübingen: Niemeyer.

Heinemann, W. (2000a). Textsorte – Textmuster – Texttyp. In K. Brinker, G. Antos, W. Heinemann & S. F. Sager (Hrsg.), *Text- und Gesprächslinguistik / Linguistics of Text and Conversation* (S. 473–488). Berlin, New York: de Gruyter

Heinemann, W. (2000b). Textsorten. Zur Diskussion um Basisklassen des Kommunizierens. Rückschau und Ausblick. In: Adamzik, K. (Hrsg.), *Textsorten. Reflexionen und Analysen* (S. 9–29). Tübingen, Stauffenburg.

Heinemann, W. (2008). Textpragmatische und kommunikative Ansätze. In N. Janich (Hrsg.), *Textlinguistik. 15 Einführungen* (S. 113–144). Tübingen: Narr.

Heinemann, W. (2009). Stilistische Phänomene auf der Ebene des Textes. In U. Fix, A. Gardt & J. Knape (Hrsg.), Rhetorik und Stilistik/Rhetoric and Stylistics. *Ein internationales Handbuch historischer und systematischer Forschung/An International Handbook of Historical and Systematic Research* (S. 1610–1630). Berlin, New York: de Gruyter.

Heinemann, W. & Viehweger, D. (1991). *Textlinguistik. Eine Einführung*. Tübingen: Niemeyer.

Helbig, G. (1975). Zu den Problemen der linguistischen Beschreibungen des Dialogs im Deutschen. *Deutsch als Fremdsprache 12*, 65–80.

Heller, M. F. (1999). *Reading-writing connections: From theory to practice*. Mahwah, NJ: Erlbaum.

Helmke, A. (2003). *Unterrichtsqualität.* Seelze: Kallmeyer.

Herder, J. G. (1786). Die Schrift. In J. G. Herder, *Zerstreute Blätter* (Zweite Sammlung) (S. 59). Gotha: Carl Wilhelm Ettinger.

Hochhaus, S. (2004). *Der verständliche Text. Perspektiven auf die Textoptimierung.* Magisterarbeit, Ruhr-Universität Bochum.

Hofen, N. (1980). *Messen und Beurteilen sprachlich-produktiver Leistungen im Deutschaufsatz: ein empirischer Beitrag zum differenzierten Erfassen schriftlicher Sprachhandlungs-Kompetenz im vierten Schuljahr.* Dissertation, Universität Mannheim.

Hoffmann, L. [Lothar] (1998a). Fachsprachen als Subsprachen. In L. Hoffmann, H. Kalverkämper & H. E. Wiegand (Hrsg.), *Fachsprachen: ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft* (S. 189–198). Berlin, New York: de Gruyter.

Hoffmann, L. [Lothar] (1998b). Fachtextsorten der Institutionensprachen III: Verträge. In L. Hoffmann, H. Kalverkämper & H. E. Wiegand (Hrsg.), *Fachsprachen: ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft* (S. 533–539). Berlin & New York: de Gruyter.

Hoffmann, L. [Ludger] (Hrsg.) (1996). *Sprachwissenschaft: ein Reader*. Berlin: de Gruyter.

Hoffmann, L. [Ludger] (2000). Thema, Themenentfaltung, Makrostruktur. In K. Brinker & G. Antos (Hrsg.), *Text- und Gesprächslinguistik – ein internationales Handbuch zeitgenössischer Forschung.* 1. Halbband (S. 344–356). Berlin, New York: de Gruyter.

Hofmeister, W. (2005). Erläuterung der Klassifikationsmatrix zum ULME-Kompetenzstufenmodell. *bwp@–Berufs-und Wirtschaftspädagogik online*, (8).

Hood, S. B. (2009). Validity in psychological testing and scientific realism. *Theory & Psychology*, *19*(4), 451–473.

Hornke, L. F., Amelang, M., Kersting, M., Birbaumer, N., Frey, D., Kuhl, J., Schneider, W. & Schwarzer, R. (Eds.) (2011). *Themenbereich B: Methodologie und Methoden / Psychologische Diagnostik / Persönlichkeitsdiagnostik* (Vol. 2). Göttingen: Hogrefe.

Houston, R. A. (2011). Literacy. In Europäische Geschichte Online (EGO). Letzter Zugriff am 30.03.2015 unter http://ieg-ego.eu/en/threads/backgrounds/literacy/robert-a-houston-literacy.

Howell, D. C. (2002). *Statistical methods for psychology* (5th edition). Pacific Grove, CA: Wadsworth.

Hox, J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar & M. Schader (Eds.), *Classification, data analysis and data highways* (147–154). New York: Springer.

Hox, J. (2002). *Multilevel Analysis: Techniques and Applications.* Mahwah, NJ: Erlbaum.

Hu, L. T. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Hubert, M. D. (2008). *The relationship between writing and speaking in the U.S. Foreign Language classroom*. Dissertation, Purdue University, West Lafayette.

Hubert, M. D. (2013). The Development of Speaking and Writing Proficiencies in the Spanish Language Classroom: A Case Study. *Foreign Language Annals 46*(1), 1–8.

Hubertus, P. (2013). Trend: Nationale Strategie zur Alphabetisierung und Grundbildung in Deutschland. *Vierteljahresschrift für Heilpädagogik und ihre Nachbargebiete*, (2), 140–142.

Hug, M. (2001). *Aspekte zeitsprachlicher Entwicklung in Schülertexten. Eine Untersuchung im 3., 5. und 7. Schuljahr*. Frankfurt am Main et al.: Lang.

Hussy, W., Schreier, M. & Echterhoff, G. (2009). *Forschungsmethoden in Psychologie und Sozialwissenschaften*. Heidelberg: Springer.

Imhof, M. (2003). *Zuhören: psychologische Aspekte auditiver Informationsverarbeitung*. Göttingen: Vandenhoeck & Ruprecht.

Ingenkamp, K.-H. (1971). *Die Fragwürdigkeit der Zensurengebung.* Weinheim: Beltz.

IQB – Institut zur Qualitätsentwicklung im Bildungswesen (2014). *Kompetenzstufenmodelle zu den Bildungsstandards im Kompetenzbereich Schreiben, Teilbereich freies Schreiben für den Mittleren Schulabschluss.* Letzter Zugriff am 30.03.2015 unter https://www.iqb.hu-berlin.de/bista/ksm/KSM_Schreiben_MS.pdf.

IQB – Institut zur Qualitätsentwicklung im Bildungswesen (2015). *Kompetenzstufenmodelle.* Letzter Zugriff am 12.07.2015 unter http://www.iqb.hu-berlin.de/bista/ksm.

Isenberg, H. (1983). Grundfragen der Textypologie. In: F. Daneš & D. Viehweger (Hrsg.), *Ebenen der Textstruktur (= Linguistische Studien Reihe A Arbeitsberichte 112)* (S. 303–342). Berlin: Akademie der Wissenschaften der DDR, Zentralinstitut für Sprachwissenschaft.

Johnson-Laird, P. N. (1983). *Mental models: Toward a cognitive science of language, inference, and consciousness.* Cambridge: Cambridge University Press.

Jonkisz, E., Moosbrugger, H. & Brandt, H. (2012). Planung und Entwicklung von Tests und Fragebogen. In H.Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (27–74). Berlin, Heidelberg: Springer.

Jude, N. (2008). *Zur Struktur von Sprachkompetenz*. Dissertation, Goethe-Universität Frankfurt am Main.

Jurecka, A. (2010). *Zum Zusammenhang von Differentiellen Item Funktionen und Testkultur.* Dissertation, Goethe-Universität Frankfurt am Main.

Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin*, *112*, 527–535.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, *21*(1), 31–41.

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th edition, pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane, M. T. (2013a). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50*(1), 1–73.

Kane, M. T. (2013b). Validation as a Pragmatic, Scientific Activity. *Journal of Educational Measurement*, *50*(1), 115–122.

Kauertz, A. (2007). *Schwierigkeitserzeugende Merkmale physikalischer Testaufgaben.* Berlin: Logos.

Kegley, P. H. (1986). The effect of mode discourse on student writing performance: Implications for policy. *Educational Evaluation and Policy Analysis*, *8*(2), 147–154.

Keibel, H. (2008, 2009). *Mathematische Häufigkeitsmaße in der Korpuslinguistik: Eigenschaften und Verwendung.* Mannheim: Institut für Deutsche Sprache. Letzter Zugriff am 30.03.2015 unter http://www.ids-mannheim.de/kl/dokumente/freqMeasures.html.

Kellogg, R. T. (1987). Effects of topic knowledge on the allocation of processing time and cognitive effort to writing processes. *Memory & Cognition*, *15*(3), 256–266.

Kempen, G. & Hoenkamp, E. (1987). An incremental procedural grammar for sentence formulation. *Cognitive science*, *11*(2), 201–258.

Kercher, J. (2010). Zur Messung der Verständlichkeit deutscher Spitzenpolitiker anhand quantitativer Textmerkmale. In T. Faas, K. Arzheimer & S. Roßteutscher (Hrsg.), *Information–Wahrnehmung–Emotion. Politische Psychologie in der Wahl und Einstellungsforschung* (S. 97–121). Wiesbaden: VS Verlag für Sozialwissenschaften.

Kessel, K. & Reimann, S. (2012). *Basiswissen Deutsche Gegenwartssprache: Eine Einführung* (4. Auflage). Stuttgart: Francke.

Kingston, N. M., Kahl, S. R., Sweeney, K. P. & Bay, L. (2001). Setting performance standards using the Body of Work method. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 219–248). Mahwah, NJ: Erlbaum.

Kintsch, W. (1974). *The representation of meaning in memory.* Hillsdale, NJ: Erlbaum.

Kintsch, W. (1982). *Gedächtnis und Kognition.* Berlin: Springer.

Kintsch, W. & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological review*, *85*(5), 363.

Klann-Delius, G. (2008). Modelle des kindlichen Wortschatzerwerbs. *Spektrum Patholinguistik I*, 1–18.

Klare, G. (1963). *The measurement of readability*. Ames: The Iowa State University Press.

Klein, W. (2007). Mechanismen des Erst-und Zweitspracherwerbs. *Sprache Stimme Gehör*, *31*, 138–143.

Klein, W. & Dimroth, C. (2009). Untutored second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *The new handbook of second language acquisition* (pp. 503–522). Bingley: Emerald.

Klieme, E. (2006). *Zusammenfassung zentraler Ergebnisse der DESI-Studie.* Frankfurt am Main: Deutsches Institut für Internationale Pädagogische Forschung (dipf).

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Ternorth, H.-E. & Vollmer, H. J. (2007). *Zur Entwicklung nationaler Bildungsstandards. Expertise*. Bonn: Bundesministerium für Bildung und Forschung (BMBF).

Klieme, E., Eichler, W., Helmke, A., Lehmann, R. H., Nold, G., Rolff, H. G., Schröder, K., Thomé, G. & Willenberg, H. (2006). *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Zentrale Befunde der Studie Deutsch-Englisch-Schülerleistungen-International (DESI). Eine Studie im Auftrag der Kultusminister der Länder in der Bundesrepublik Deutschland*. Frankfurt am Main: Deutsches Institut für Internationale Pädagogische Forschung (dipf).

Klieme, E. & Hartig, J. (2008). Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. In M. Prenzel, I. Gogolin & H. H. Krüger (Hrsg.), *Kompetenzdiagnostik* (S. 11–29). Wiesbaden: Springer VS.

Klieme, E. & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik*, *54*(2), 222–237.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd edition). New York: Guilford.

Klug, A. (2007). *Knowledgebase – Kompetenz.* Letzter Zugriff am 30.06.2015 unter http://klug-md.de/Wissen/Kompetenz.htm.

KMK (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland) (1997). *Grundsätzliche Überlegungen zu Leistungsvergleichen innerhalb der Bundesrepublik Deutschland – Konstanzer Beschluss.* Beschluss der Kultusministerkonferenz vom 24.10.1997. Letzter Zugriff am 30.03.2015 unter http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/1997/1997_10_24-Konstanzer-Beschluss.pdf.

KMK (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland) (2004). *Bildungsstandards im Fach Deutsch für den Mittleren Schulabschluss. Beschluss vom 4.12.2003*. München: Luchterhand.

KMK (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland) (2005a). *Bildungsstandards im Fach Deutsch für den Hauptschulabschluss. Beschluss vom 15.10.2004.* München: Luchterhand.

KMK (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland) (2005b). *Bildungsstandards der Kultusministerkonferenz – Erläuterungen zur Konzeption und Entwicklung*. München: Luchterhand.

Knapp, P. & Watkins, M. (2005). *Genre, text, grammar: Technologies for teaching and assessing writing*. Sydney: UNSW Press.

Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik* (Vol. 31). Bochum: Brockmeyer.

Köhler, R. & Altmann, G. (1986). Synergetische Aspekte der Linguistik. *Zeitschrift. für Sprachwissenschaft 5, 253-265.*

Köller, O. (2008). Bildungsstandards in Deutschland: Implikationen für die Qualitätssicherung und Unterrichtsqualität. In M. Meyer, M. Prenzel & S. Hellekamps (Hrsg.), *Perspektiven der Didaktik* (S. 47–59). Wiesbaden: VS Verlag für Sozialwissenschaften.

Köller, O. (2010). Bildungsstandards. In R. Tippelt & B. Schmidt (Hrsg.), *Handbuch der Bildungsforschung* (3. Auflage, S. 529–548). Wiesbaden: VS Verlag für Sozialwissenschaften.

Köller, O., Knigge, M. & Tesch, B. (Hrsg.) (2010). *Sprachliche Kompetenzen im Ländervergleich*. Münster: Waxmann

Köster, J. (2005). Wodurch wird ein Test schwierig? Ein Text für die Fachkonferenz. *Deutschunterricht, 58*(5)*,* 34–39.

Krause, T. & Urban, D. (2013). *Panelanalyse mit Mehrebenenmodellen. Eine anwendungsorientierte Einführung*. Schriftenreihe des Instituts für Sozialwissenschaften der Universität Stuttgart, SISS No. 1/2013.

Kreft, I. G. G. (1996). *Are multilevel techniques necessary? An overview, including simulation studies.* Unpublished manuscript, California State University, Los Angeles.

Kreiner, D. S., Schnakenberg, S. D., Green, A. G., Costello, M. J. & McClin, A. F. (2002). Effects of spelling errors on the perception of writers. *Journal of General Psychology*, *129*, 5–17.

Kromrey, J. D. & Hogarty, K. Y. (1998). Analysis options for testing group differences on ordered categorical variables: An empirical investigation of type I error control and statistical power. *Multiple linear regression viewpoints, 25*(1), 70–82.

Kruskal, W. H. & Wallis, W. A. (1953). Errata: Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *48*(264), 907–911.

Kubinger, K. D. (2009). *Psychologische Diagnostik: Theorie und Praxis psychologischen Diagnostizierens* (2. Auflage). Göttingen: Hogrefe.

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., Prenzel, M, Schiefele, U., Schneider, W., Stanat, P., Tilmann, K.-J. & Weiß, M. (2001). *PISA 2000: Dokumentation der Erhebungsinstrumente.* Berlin: Max-Planck-Institut für Bildungsforschung.

Kürschner, C. & Schnotz, W. (2008). Das Verhältnis gesprochener und geschriebener Sprache bei der Konstruktion mentaler Repräsentationen. *Psychologische Rundschau*, *59*(3), 139–149.

Lado, R. (1957). *Linguistics across cultures: Applied linguistics for language teachers*. Ann Arbor: University of Michigan Press.

Landerl, K. & Wimmer, H. (2000). Deficits in phoneme segmentation are not the core problem of dyslexia: Evidence from German and English children. *Applied psycholinguistics*, *21*(2), 243–262.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Langer, J. & Flihan, S. (2000). Writing and reading relationships: Constructive tasks. In R. Indrisano & J. R. Squire (Eds.), *Writing: Research / Theory / Practice* (pp. 112–139). Newark, DE: International Reading Association.

Langer, W. (2010). Mehrebenenanalyse mit Querschnittsdaten. In C. Wolf & H. Best (Hrsg.), *Handbuch der sozialwissenschaftlichen Datenanalyse* (S. 741–774). Wiesbaden: VS Verlag für Sozialwissenschaften.

Langsrud, Ø. (2003). ANOVA for unbalanced data: Use Type II instead of Type III sums of squares. *Statistics and Computing*, *13*(2), 163–167.

LeBreton, J. M. & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*, 815–852.

Lees-Haley, P. R. (1996). Alice in Validityland, or the dangerous consequences of consequential validity. *American Psychologist, 51*, 981–983.

Lehmann, R. H. (1987). The Scoring of Students Compositions. In W. Hartmann & R. H. Lehmann (Eds.), *The Hamburg Study of Achievement in Written Composition. National Report for the IEA International Study of Achievement Written Communication.* Hamburg: University of Hamburg.

Lehmann, R. H. (1990). Aufsatzbeurteilung – Forschungsstand und empirische Daten. In K. Ingenkamp, K. & R. Jäger (Hrsg.), *Tests und Trends: Jahrbuch der Pädagogischen Diagnostik* (Band 8, S. 64–94). Weinheim, Basel: Beltz.

Lehmann, R. H. (1994). Research on National and International Writing Assessments: Contributions from the Hamburg Study of Achievement in Written Composition. In R. Ansorge (Hrsg.), *Schlaglichter der Forschung. Zum 75. Jahrestag der Universität Hamburg* (S. 173–184). Berlin, Hamburg: Reimer.

Lehmann, R. H. & Hartmann, W. (1987). *The Hamburg study of achievement in written composition. National report for the IEA international study of achievement in written composition.* Hamburg.

Lehmann, R. H., Hunger, S., Ivanov, S., Gänsfuß, R. & Hoffmann, E. (2004). *Aspekte der Lernausgangslage und der Lernentwicklung – Klassenstufe 11. Ergebnisse einer Längsschnittsuntersuchung in Hamburg.* Hamburg: Behörde für Bildung und Sport.

Lehmann, R. H. & Peek, R. (1997). *Aspekte der Lernausgangslage von Schülerinnen und Schülern der fünften Klassen an Hamburger Schulen. Bericht über die Untersuchung im September 1996*. Hamburg: Behörde für Bildung und Sport.

Lehmann, R. H., Peek, R., Gänsfuß, R. & Husfeldt, V. (2002). *Aspekte der Lernausgangslage und der Lernentwicklung – Klassenstufe 9. Ergebnisse einer Längsschnittsuntersuchung in Hamburg.* Hamburg: Behörde für Bildung und Sport.

Leppert, U. (2010). *Ich hab eine Eins! Und Du?: Von der Notenlüge zur Praxis einer besseren Lernkultur*. BoD-Books on Demand.

Levelt, W. J. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

Liao, C., Qu, Y. & Morgan, R. (2010). *The Relationships of Test Scores Measured by the TOEIC Listening and Reading Test and TOEIC Speaking and Writing Tests*. Letzter Zugriff am 30.03.2015 unter http://www.ets.org/Media/Research/pdf/TC-10-13.pdf.

Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Auflage). Weinheim: Beltz.

Linn, R. L., Klein, S. P. & Hart, F. M. (1972). The nature and correlates of law school essay grades. *Educational and Psychological Measurement*, *32*, 267–279.

Loftus, G. R. (1971). Comparison of recognition and recall in a continuous memory task. *Journal of experimental psychology*, *91*(2), 220–226.

Ludwig, O. (1988). *Der Schulaufsatz. Seine Geschichte in Deutschland.* Berlin, New York: de Gruyter.

Ludwig, O. (2003). Entwicklung schulischer Schreibdidaktik in Deutschland und ihr Bezug zu akademischen Schreiben. In K. Ehlich und A. Steets (Hrsg.), *Wissenschaftlich schreiben – lehren und lernen* (S. 235–250). Berlin, New York: de Gruyter.

Ludwig, O. (2006). Geschichte der Didaktik des Texteschreibens. In U. Bredel, H. Günther, P. Klotz, P. Ossner & G. Siebert-Ott (Hrsg.), *Didaktik der deutschen Sprache. Ein Handbuch* (1. Halbband, 2. Auflage, S. 171–177). Paderborn: Schöningh.

Maas, C. J. M. & Hox, J. J. (2004). Robustness issues in multi-level regression analysis. *Statistica Neerlandica, 58*, 127–137.

Macbeth, G., Razumiejczyk, E. & Ledesma, R. D. (2011). Cliff's Delta Calculator: A non-parametric effect size program for two groups of observations. *Universitas Psychologica*, *10*(2), 545–555.

Maehler, D. & Schmidt-Denter, U. (2013). *Migrationsforschung in Deutschland: Leitfaden und Messinstrumente zur Erfassung psychologischer Konstrukte.* Wiesbaden: Springer.

Maia, M. (2008). Reading and Listening to Garden-Path PP Sentences in Brazilian Portuguese. *Trabalho aceito para apresentação no Speech Prosody*, 6–9.

Mann, H. B. & Whitney, D. R. (1947). On a test whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics 18*, 50–60.

Marquard, O. (1981). *Abschied vom Prinzipiellen: Philosophische Studien*. Stuttgart: Reclam.

Marshall, J. C. (1967). Composition errors and essay examination grades re-examined. *American Educational Research Journal*, *4*, 375–385.

Marshall, J. C. & Powers, J. C. (1969). Writing neatness, composition errors, and essay grades. *Journal of Educational Measurement*, *6*, 97–101.

Mathesius, V. (1929). Zur Satzperspektive im modernen Englisch. *Archiv für das Studium der neueren Sprachen und Literaturen 84*, 202–210.

McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review*, *8*(3), 299–325.

McKoon, G. & Ratcliff, R. (1992). Inference during reading. *Psychological review*, *99*(3), 440.

Meisel, J. M., Clahsen, H. & Pienemann, M. (1981). On determining developmental stages in natural second language acquisition. *Studies in Second Language Acquisition 3*(2),109–135.

Merkens, H. (2006). Bildungsforschung und Erziehungswissenschaft. In H. Merkens (Hrsg.), *Erziehungswissenschaft und Bildungsforschung* (S. 9–20). Wiesbaden: VS Verlag für Sozialwissenschaften.

Merz-Grötsch, J. (2001). *Schreiben als System. Band 2: Die Wirklichkeit aus Schülersicht.* Freiburg im Breisgau: Fillibach.

Merz-Grötsch, J. (2005). *Schreiben als System. Band 1: Schreibforschung und Schreibdidaktik. Ein Überblick* (2. Auflage). Freiburg: Fillibach.

Merz-Grötsch, J. (2010). *Texte schreiben lernen: Grundlagen, Methoden, Unterrichtsvorschläge*. Seelze: Klett Kallmeyer.

Messelken, H. (1971). *Empirische Sprachdidaktik*. Heidelberg: Quelle & Meyer.

Messick, S. (1980). Test validity and the ethics of assessment. *American psychologist*, *35*(11), 1012–1027.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher*, *10*(9), 9–20.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd edition, pp. 13–103). New York, NY: American Council on Education and Macmillan.

Messick, S. (1990). *Validity of test interpretation and use.* (Report No. ETS-RR-90-11). Princeton, NJ: Educational Testing Service.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*, 13–23.

Messick, S. (1995a). Standards of validity and the validity of standards in performance assessment. *Educational measurement: Issues and practice*, *14*(4), 5–8.

Messick, S. (1995b). Validity of psychological assessment: Validation of inferences from persons' re-sponses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749.

Messick, S. (1996). Validity and washback in language testing. *ETS Research Report Series*, *1996*(1), 1–18.

Meurers, D. (2014). *Zur automatischen Analyse der Lesbarkeit von Texten und Sätzen.* Präsentation: Deutsch 3.0 Workshop "Text als Werkstück – Wege zu einer computergestützten Überarbeitung von deutschen Texten", Deutsches Institut für Internationale Pädagogische Forschung, Frankfurt am Main, 7. Juli 2014.

Mihm, A. (1973). Sprachstatistische Kriterien zur Tauglichkeit von Lesebüchern. *Linguistik und Didaktik 4*, 117–127.

Mikk, J. (1995). Methods for determining optimal readability of texts. *Journal of Quantitative Linguistics*, *2*(2), 125–132.

Mikk, J. (2000). *Textbook Research* and *Writing*. Frankfurt am Main et al.: Lang.

Mikk, J. & Elts, J. (1999). A reading comprehension formula of reader and text characteristics. *Journal of Quantitative Linguistics*, *6*(3), 214–221.

Mislevy, R. J., Beaton, A. E., Kaplan, B. & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133–161.

Molitor-Lübbert, S. (1996). Schreiben als mentaler und sprachlicher Prozess. In H. Günther & O. Ludwig (Hrsg.), *Schrift und Schriftlichkeit. Ein interdisziplinäres Handbuch internationaler Forschung* (S. 1005–1027). Berlin, New York: de Gruyter.

Motsch, W. & Viehweger, D. (1981). Sprachhandlung, Satz und Text. In I. Rosengren (Hrsg.), *Sprache und Pragmatik. Lunder Symposium 1980* (S. 125–154). Malmö: Gleerup.

Murphy, K. R. & Davidshofer, C. O. (2001). *Psychological testing: principles and applications* (5th edition). Upper Saddle River, NJ: Prentice-Hall.

Muthén, L. K. & Muthén, B. O. (1998). *Mplus. Statistical analyses with latent variables.* Los Angeles, CA: Muthén & Muthén.

Muthén, L. K. & Muthén, B. O. (2008). *Mplus (Version 5.1).* Los Angeles, CA: Muthén & Muthén.

Myers, L. & Sirois, M. J. (2006). Spearman correlation coefficients, differences between. *Wiley StatsRef: Statistics Reference Online*.

NAEP (1999): Grennwald, E. A., Persky, H. R., Campbell, J. R. & Mazzeo, J. (1999). *NAEP 1998. Report Card for the Nation and the states.* Office of Educational Research and Improvement. National Center for Education Statistics, Washington D.C.: U.S. Department of Education.

NAEP (2001): Loomis, S. C. & Bourque, M. L. (Eds.) (2001). *National Assessment of Educational Progress Achievement Levels 1992–1998 for Writing.* Washington D.C.: U.S. Department of Education.

NAEP (2003): Persky, H. R., Daane, M. C. & Ying, J. (2003). *The Nation's Report Card: Writing 2002.* National Center for Education Statistics, Institute of Education Sciences, Washington D.C.: U.S. Department of Education.

NAEP (2008): Salahu-Din, D., Persky, H. & Miller, J. (2008). *The Nation's Report Card: Writing 2007.* Washington D.C.: U.S. Department of Education.

NAEP (2011a): National Assessment Governing Board (2011). *Developing Achievement Levels on the National Assessment of Educational Progress for Writing Grades 8 and 12 in 2011 and Grade 4 in 2013.* NAEP Writing ALS Design Document.

NAEP (2011b): National Assessment Governing Board (2011). *Writing Framework for the 2011 National Assessment of Educational Progress.* U.S. Department of Education, Washington, D.C.

NAEP (2012): National Center for Education Statistics (2012). *The Nation's Report Card: Writing 2011.* Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

Necknig, A. T. (2007). *Schreibkonferenz vs. traditionelle Aufsatzdidaktik. Eine empirische Untersuchung.* Dissertation, Universität Koblenz-Landau.

Neumann, A. (2007). *Briefe schreiben in Klasse 9 und 11. Beurteilungskriterien, Messungen, Textstrukturen und Schülerleistungen.* Münster, New York, München, Berlin: Waxmann.

Neumann, A. (2012). Advantages and Disadvantages of Different Text Coding Procedures for Research and Practice in a School Context. In E. van Steendam, M. Tillema, G. Rijlaarsdam & H. van den Bergh (Eds.), *Measuring writing: recent insights into theory, methodology and practices* (Vol. 27, pp. 33–54). Leiden: Brill.

Neumann, A. (2014). Großuntersuchungen zur Schreibleistungsmessung. In H. Feilke & T. Pohl (Hrsg.), *Schriftlicher Sprachgebrauch – Texte verfassen* (S. 514–531). Baltmannsweiler: Schneider Hohengehren.

Neumann, O., van der Heijden, A. H. C. & Allport, D. A. (1986). Visual selective attention: Introductory remarks. *Psychological Research*, *48*(4), 185–188.

Newton, P. & Shaw, S. (2012). *The meaning of validity: consensus, what consensus?* Presentation at The Validity Symposia, University of Cambridge, February 29th, 2012.

Newton, P. & Shaw, S. (2014). *Validity in Educational and Psychological Assessment*. London: SAGE.

Nezlek, J. B., Schröder-Abé, M. & Schütz, A. (2006). Mehrebenenanalysen in der psychologischen Forschung. *Psychologische Rundschau*, *57*(4), 213–223.

Niznikiewicz, M. & Squires, N. K. (1996). Phonological processing and the role of strategy in silent reading: behavioral and electrophysiological evidence. *Brain and Language*, *52,* 342–364.

Nold, G. & Rossa, H. (2007). Leseverstehen. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen: Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistung International)* (S. 197–211). Weinheim: Beltz.

Nunan, D. & Koepke, K. (1995). Task difficulty from the learner's perspective: Perceptions and reality. *Hong Kong Papers in Linguistics and language Teaching, 18*(1), 1–12.

Nussbaumer, M. (1991). *Was Texte sind und wie sie sein sollen: Ansätze zu einer sprachwissenschaftlichen Begründung eines Kriterienrasters zur Beurteilung von schriftlichen Schülertexten*. Tübingen: Niemeyer.

Nutz, M. (2006). Beurteilung sprachlicher Leistungen. In U. Bredel, H. Günther, P. Klotz, J. Ossner & G. Siebert-Ott (Hrsg.), *Didaktik der deutschen Sprache. Ein Handbuch* (2. Teilband, 2. Auflage, S. 924–940). Paderborn: Schöningh.

OECD (2004). *Lernen für die Welt von morgen. Erste Ergebnisse von PISA2003*. Paris: OECD.

Ohlhus, S. (2014). Schrifliches Erzählen. In H. Feilke & T. Pohl (Hrsg.), *Schriftlicher Sprachgebrauch – Texte verfassen* (S. 216–232). Baltmannsweiler: Schneider Hohengehren.

Olinghouse, N. G., Santangelo, T. & Wilson, J. (2012). Examining the Validity of Single-Occasion, Single-Genre, Holistically Scored Writing Assessments. In E. van Steendam, M. Tillema, G. Rijlaarsdam & H. van den Bergh (Eds.), *Measuring writing. Recent*

insights into theory, methodology and practices (pp. 55–82). Leiden: Brill Academic Publishers.

Ossner, J. (1995). Prozessorientierte Schreibdidaktik in Lehrplänen. In J. Baurmann & R. Weingarten (Hrsg.), *Schreiben. Prozesse, Prozeduren und Produkte* (S. 29–50). Opladen: Westdeutscher Verlag.

Ossner, J. (2006). *Sprachdidaktik Deutsch*. Paderborn: Schöningh.

Ossner, J. (2014). Schriftliches Beschreiben. In H. Feilke & T. Pohl (Hrsg.), *Schriftlicher Sprachgebrauch – Texte verfassen* (S. 252–269). Baltmannsweiler: Schneider Hohengehren.

Pant, H. A., Böhme, K. & Köller, O. (2012). Das Kompetenzkonzept der Bildungsstandards und die Entwicklung von Kompetenzstufenmodellen. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik* (S. 49–55). Münster: Waxmann.

Pant, H. A., Tiffin-Richards, S. P. & Köller, O. (2010). Standard-Setting für Kompetenztests im Large-Scale-Assessment. Projekt Standardsetting. In E. Klieme, D. Leutner & M. Kenk (Hrsg), *Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunkt-programms und Perspektiven des Forschungsansatzes* (S. 175–188). Weinheim, Basel: Beltz.

Parodi, G. (2007). Reading-writing connections: Discourse-oriented research. *Reading and Writing, 20*(3), 225–250.

Perfetti, C. A. (1992). The representation problem in reading acquisition. In P. B. Gough, L. C. Ehri & R. Treiman (Eds.), *Reading acquisition* (pp. 145–174). Hillsdale, NJ: Erlbaum.

Perfetti, C. A. (1999). Comprehending written language: A blueprint of the reader. In C. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 167–208). Oxford: Oxford University Press.

Perkuhn, R., Keibel, H. & Kupietz, M. (2012). *Korpuslinguistik – Begleitmaterialien. Berechnung von Häufigkeitsklassen.* Letzter Zugriff am 30.03.2015 unter http://corpora.ids-mannheim.de/libac/hk.shtml.

Pfister, B & Kaufmann, T. (2008). *Sprachverarbeitung – Grundlagen* und *Methoden* der *Sprachsynthese* und *Spracherkennung*. Berlin, Heidelberg: Springer.

PISA-Konsortium Deutschland (Hrsg.) *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs.* Waxmann, 2004.

Pohl, T. (2007). *Studien zur Ontogenese wissenschaftlichen Schreibens*. Tübingen: Niemeyer.

Pohl, T. (2014). Schriftliches Argumentieren. In H. Feilke & T. Pohl (Hrsg.), *Schriftlicher Sprachgebrauch – Texte verfassen* (S. 287–315). Baltmannsweiler: Schneider Hohengehren.

Polenz, P. v. (1988). *Deutsche Satzsemantik. Grundbegriffe des Zwischenden-Zeilen-Lesens* (2. Auflage). Berlin, New York: de Gruyter.

Pollatsek, A., Ashby, J. & Clifton Jr, C. (2012). *Psychology of reading* (2nd edition). New York: Psychology Press.

Powers, D. E., Kim, H.-J., Yu, F., Weng, V. Z. & VanWinkle, W. (2009). *The TOEIC Speaking and Writing Tests: Relations to Test-Taker Perceptions of Proficiency in English.* Letzter Zugriff am 30.03.2015 unter http://www.ets.org/Media/Research/pdf/RR-09-18.pdf.

Prabhu, N. S. (1987). *Second language pedagogy*. Oxford: University Press.

Prater, D. & Padia, W. (1983). Effects of modes of discourse on writing performance in grades four and six. *Research in the Teaching of English*, *17*, 127–134.

Prenzel, M. (Hrsg.) (2007). *PISA 2006: Die Ergebnisse der dritten internationalen Vergleichsstudie*. Münster, New York: Waxmann.

Prose, F. (2006). *Reading Like A Writer: A Guide for People Who Love Books and for Those Who Want to Write Them.* New York: HarperCollins.

Püschel, U. (2000). Text und Stil. In K. Brinker, G. Antos, W. Heinemann & S. F. Sager (Hrsg.), *Text- und Gesprächslinguistik / Linguistics of Text and Conversation* (S. 473–488). Berlin, New York: de Gruyter.

Quasthoff, U. M. (1980). *Erzählen in Gesprächen: linguistische Untersuchungen zu Strukturen und Funktionen am Beispiel einer Kommunikationsform des Alltags*. Tübingen: Narr.

Quasthoff, U. M. (1993). Dabeisein durch Sprache: zur Rolle der Perspektive beim konversationellen Erzählen. In P. Cansius & M. Gerlach (Hrsg.), *Perspektivität in Sprache und Text* (2. Auflage, S.129–151). Bochum: Brockmeyer.

Quellmalz, E. S., Capell, F. J. & Chou, C. P. (1982). Effects of discourse and response mode on the measurement of writing competence. *Journal of Educational Measurement*, *19*(4), 241–258.

Raghunathan, T. E., Rosenthal, R. & Rubin, D.B. (1996). Comparing correlated but nonoverlapping correlations. *Psychological Methods, 1*(2), 178–183.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Kopenhagen: Paedagogiske Institut.

Ratcliff, R. & McKoon, G. (1978). Priming in item recognition: Evidence for the propositional structure of sentences. *Journal of verbal learning and verbal behavior*, *17*(4), 403–417.

Rayner, K., Carlson, M. & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of verbal learning and verbal behavior*, *22*(3), 358–374.

Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice*, *17*(2), 13–16.

Rehbein, J. (1983). Beschreiben, berichten und erzählen. In K. Ehlich (Hrsg.), *Erzählen in der Schule* (S. 67–124). Tübingen: Narr.

Rehbein, J. (1988). Ausgewählte Aspekte der Pragmatik. In L. Hoffmann, (Hrsg.) (1996), *Sprachwissenschaft: ein Reader* (S. 106–131). Berlin, New York: de Gruyter.

Rehder, A. (2011). *Was ist eigentlich »Schreibkompetenz«?* GRIN.

Rheinberg, F., Vollmeyer, R. & Burns, B. D. (2001). FAM : Ein Fragebogen zur Erfassung aktueller Motivation [QCM : A questionnaire to assess current motivation in learning situations]. *Diagnostica*, *47*, 1–17.

Richter, B. (2008). *Didaktische Konzepte zur Förderung der Schreibkompetenz in der Sekundarstufe I.* Essen: LINSE. Letzter Zugriff am 30.03.2015 unter http://www.linse.uni-due.de/linse/esel/arbeiten/schreibkompetenzfoerderung.pdf

Rickheit, G., Herrmann, T. & Deutsch, W. (Hrsg.) (2003). *Psycholinguistik/ Psycholinguistics: Ein internationales Handbuch/An International Handbook* (Vol. 24). Berlin, New York: de Gruyter.

Rickheit, G., Sichelschmidt, L. & Strohner, H. (2004). *Psycholinguistik: Die Wissenschaft vom sprachlichen Verhalten und Erleben.* Tübingen: Stauffenburg

Rickheit, G. & Strohner, H. (1993). *Grundlagen der kognitiven Sprachverarbeitung: Modelle, Methoden, Ergebnisse*. Tübingen: Francke.

Riesel, E. & Schendels, E. (1975). *Deutsche Stilistik.* Moskau: Hochschulverlag.

Risel, H. (2011). *Richtig gute Aufsätze schreiben: didaktische Grundlagen und vielfältige Arbeitsblätter zu den verschiedenen Textsorten; 3./4. Klasse*. Augsburg: Brigg Pädagogik.

Ritter, M. (2008). *Wege ins Schreiben. Eine Studie zur Schreibdidaktik in der Grundschule.* Baltmannsweiler: Schneider Hohengehren.

Roelcke, T. (Hrsg./Ed.) (2003). *Variationstypologie/Variation Typology: Ein sprach-typologisches Handbuch der europäischen Sprachen in Geschichte und Gegenwart/ A Typological Handbook of European Languages*. New York, Berlin: de Gruyter.

Roick, T. (2008). Standardisierte Schulleistungstests. In W. Schneider & M. Hasselhorn, (Hrsg.), *Handbuch der Pädagogischen Psychologie*, *10*, (S. 271–281). Göttingen: Hogrefe.

Rosebrock, C. (2012). *Kriterien der Textbeurteilung für die Aufgabenkonstruktion.* Präsentation: Aufgabenentwicklertagung VERA-8, Universität zu Köln, 02.10.2012.

Rost, D. H. & Hartmann, A. (1992). Lesen, Hören, Verstehen. *Zeitschrift für Psychologie, 200*, 345–361.

Ruch, G. M. (1924). *The improvement of the written examination*. Chicago, IL: Scott, Foresman and Company.

Ruland, A., Willmes, K. & Günther, T. (2012). Zusammenhang zwischen Aufmerksamkeitsdefiziten und Lese-Rechtschreibschwäche. *Kindheit und Entwicklung*, *21*(1), 57–63.

Rupp, G. (1986). „In der Anarchie der Sprache gar schöne Ordnung“ sehen. Ästhetische Schulung in den Stilübungen im Literaturunterricht des 18. und 19. Jahrhunderts. In H. U. Gumbrecht & K. L. Pfeiffer (Hrsg.), *Stil – Geschichten und Funktionen eines kulturwissenschaftlichen Diskurselements* (S. 394–409). Frankfurt am Main: Suhrkamp.

Russell, M. & Tao, W. (2004). The influence of computer-print on rater scores. *Practical Assessment, Research & Evaluation*, *9*(10), 1–14.

Sáenz, L. M. & Fuchs, L. S. (2002). Examining the reading difficulty of secondary students with learning disabilities expository versus narrative text. *Remedial and Special Education*, *23*(1), 31–41.

Sandig, B. (2006). *Textstilistik des Deutschen* (2. Auflage). Berlin, New York: de Gruyter.

Sarris, V. & Reiß, S. (2005). *Kurzer Leitfaden der Experimentalpsychologie*. München: Pearson Studium.

Sauter, H. & Pschibul, M. (1977). *Vom Aufsatzunterricht zur sprachlichen Kommunikation in der Sekundarstufe I.* Donauwörth: Auer.

Schank, G. & Schoenthal, G. (1976). *Gesprochene Sprache: eine Einführung in die Forschungsansätze und Analysemethoden.* Tübingen: Niemeyer.

Scheerens, J. & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon.

Schießl, M. (1889). *Die stilistische Entwicklungstheorie in der Volksschule. Theorie, Praxis und Methodik des Aufsatzunterrichts. Eine neue Schulstilistik für Volksschullehrer.* München: Max Kellerer’s Hofbuchhandlung.

Schmidt-Atzert, L. & Amelang, M. (2012). *Psychologische Diagnostik (Lehrbuch mit Online-Materialien* (5. Auflage). Heidelberg: Springer.

Schneider, F. & Tetling, K. (2014). Argumentierend Schreiben. In M. Becker-Mrotzek & I. Böttcher (Hrsg.), *Schreibkompetenz entwickeln und beurteilen. [Sekundarstufe I/II]* (5. Auflage). Berlin: Cornelsen.

Schnell, R., Hill, P. B. & Esser, E. (2011). *Methoden der empirischen Sozialforschung* (9. Auflage). München: Oldenbourg.

Schöler, H. (2006). Sprachleistungsmessungen. In U. Bredel, H. Günther, P. Klotz, J. Ossner & G. Siebert-Ott (Hrsg.), *Didaktik der deutschen Sprache. Ein Handbuch* (2. Teilband, 2. Auflage, S. 898–913). Paderborn: Schöningh.

Schoonen, R. (2012). The validity and generalizability of writing scores: The effect of rater, task and language. In E. van Steendam, M. Tillema, G. Rijlaarsdam & H. van den Bergh (Eds.), *Measuring writing: recent insights into theory, methodology and practices* (Vol. 27, pp. 1–22). Leiden: Brill.

Schorer, H. (1959). „Die Bedeutung Wilhelm von Humboldts und Leo Weisgerbers für den Deutschunterricht in der Volksschule". In H. Gipper (Hrsg.), *Sprache – Schlüssel zur Welt: Festschrift für Leo Weisgerber* (S. 106–122). Düsseldorf: Pädagogischer Verlag Schwann.

Schott, F. & Ghanbari, S. A. (2008). *Kompetenzdiagnostik, Kompetenzmodelle, kompetenz-orientierter Unterricht*. Münster: Waxmann.

Schroeders, U. & Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement*, *71*(5), 849–869.

Schröter, G. (1971). *Die ungerechte Aufsatzzensur.* Bochum: Kamp.

Schumann, S. & Eberle, F. (2011). Bedeutung und Verwendung schwierigkeitsbestimmender Aufgabenmerkmale für die Erfassung ökonomischer und beruflicher Kompetenzen. In U. Faßhauer, B. Fürstenau & E. Wuttke (Hrsg.), *Grundlagenforschung zum Dualen System und Kompetenzentwicklung in der Lehrerbildung* (S. 77–90). Opladen, Berlin, Farmington Hills, Mich.: Budrich.

Schuster, K. (1998). *Mündlicher Sprachgebrauch im Deutschunterricht: Denken-Sprechen-Handeln; Theorie und Praxis*. Baltmannsweiler: Schneider Hohengehren.

Schwartz, B. D. & Sprouse, R. A. (1996). L2 cognitive states and the full transfer/full access Model. *Second Language Research 12*(1), 40–72.

Schweitzer, K. (2006). *Der Schwierigkeitsgrad von Textverstehensaufgaben*. *Ein Beitrag zur Differenzierung und Präzisierung von Aufgabenbeschreibungen.* Frankfurt am Main: Lang.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press.

Seel, N. M., Pirnay-Dummer, P. & Ifenthaler, D. (2010). Quantitative Bildungsforschung. In R. Tippelt & B. Schmidt (Hrsg.), *Handbuch Bildungsforschung* (S. 551–570). Wiesbaden: VS Verlag für Sozialwissenschaften.

Seifert, J. N. (1835). Der Unterricht in der deutschen Sprache zunächst für Landschulen bearbeitet. *Magazin für Elementarschullehrer 4*, Heft 1, 83–100.

Seitz, K. (2003). Der schiefe Turm von PISA – nur die Spitze eines Eisbergs? Der PISA-Schock und der weltweite Umbau der Bildungssysteme. *ZEP: Zeitschrift für internationale Bildungsforschung und Entwicklungspädagogik*, *26*(1), 2–8.

Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591–611.

Shermis, M. D., Shneyderman, A. & Attali, Y. (2008). How important is content in the ratings of essay assessments? *Assessment in Education: Principles, Policy & Practice*, *15*(1), 91–105.

Shohamy, E. (1996). Competence and performance in language testing. In G. Bronn, K. Malmkjer & J. Williams (Eds.), *Performance and Competence in Second Language Acquisition* (pp. 136–151). Cambridge: Cambridge University Press.

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical methods in medical research*, *7*(3), 301–317.

Sieber, P. (2006). Modelle des Schreibprozesse. In U. Bredel, H. Günther, P. Klotz, G. Siebert-Ott, G. & J. Ossner (Hrsg.), *Didaktik der deutschen Sprache. Ein Handbuch* (1. Teilband, 2. Auflage, S. 208–223). Paderborn et al.: Schöningh.

Singer, M. (1990). *Psychology* of *language*: *An introduction to sentence and discourse processes*. Hillsdale, NJ: Erlbaum.

Smith, M., III (2009). *The Reading-Writing Connection*. Metametrics Position Paper. Letzter Zugriff am 30.03.2015 unter http://cdn.lexile.com/m/resources/materials/Reading-Writing_Connection.pdf.

Snijders, T. & Bosker, R. (1994). Modeled variance in two-level models. *Sociological Methods & Research*, *22*(3), 342–363.

Spencer, L. M. & Spencer, S. (1993). *Competence at work: Models for superior performance.* New York: John Wiley.

Spinner, K. H. (1980). Identitätsgewinnung als Aspekt des Aufsatzunterrichts. In K. H. Spinner (Hrsg.), *Identität und Deutschunterricht* (S. 67–80). Göttingen: Vandenhoeck & Ruprecht.

Spinner, K. H. (1993). Kreatives Schreiben. *Praxis Deutsch 119*, 17–23.

Sprouse, J. L. & Webb, J. E. (1994). *The Pygmalion Effect and Its Influence on the Grading and Gender Assignment on Spelling and Essay Assessments*. Master's Thesis, University of Virignia.

Stanat, P., Pant, H. A., Böhme, K. & Richter, D. (Hrsg.) (2012). *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik: Ergebnisse des IQB-Ländervergleichs 2011*. Münster: Waxmann.

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly, 16*, 32–71.

Stanovich, K. E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. New York: Guilford.

Statistisches Bundesamt (2011). Bildung und Kultur. Allgemeinbildende Schulen. Schuljahr 2010/2011. *Fachserie 11, Reihe 1*. Letzter Zugriff am 30.03.2015 unter http://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/Schulen/AllgemeinbildendeSchulen2110100117004.pdf.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245–251.

Steinhoff, T. (2009): Wortschatz – eine Schaltstelle für den schulischen Spracherwerb? In H. Feilke, K.-P. Kappest & C. Knoblauch (Hrsg), *Siegener Papiere zur Aneignung sprachlicher Strukturformen. Schriftenreihe der Universität Siegen,* Heft 17. Universität Siegen.

Stieglitz, R. D. (2008). *Diagnostik und Klassifikation in der Psychiatrie*. Stuttgart: Kohlhammer.

Sundre, D. L. & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology*, *29*, 6–26.

Swartz, C. W., Hooper, S. R., Montgomery, J. W., Wakely, M. B., De Kruif, R. E., Reed, M., Brown, T. T., Levine, M. D. & White, K. P. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytical scoring methods. *Educational and Psychological Measurement*, *59*(3), 492–506.

Tatham, M. & Morton, K. (2010). Two Theories of Speech Production and Perception. In J. Guendouzi, F. Loncke & M. J.Williams (Eds.), *The Handbook of Psycholinguistic and Cognitive Processes* (pp. 291–293). New York, London: Psychology Press.

Taylor, B. M. & Beach, R. W. (1984). The effects of text structure instruction on middle-grade students' comprehension and production of expository text. *Reading Research Quarterly*, 134–146.

Tent, L. (1998). Zensuren. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (S. 580–584). Weinheim: Psychologie Verlags Union.

Thorndike, E. L. (1920). A constant error in psychological rating. *Journal of Applied Psychology*, *4*, 25–29.

Tiemann, R. & Körbs, C. (2014). Die Fragebogenmethode, ein Klassiker der empirischen didaktischen Forschung. In I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (283–295). Berlin, Heidelberg: Springer.

Tinsley, H. E. A. & Weiss, D. J. (2000). Interrater reliability and agreement. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (95–124). New York: Academic Press.

Toulmin, S. (1958). *The Uses of Argument*. Cambridge: Cambridge University Press.

Treiman, R., Clifton, C. Jr., Meyer, A. S. & Wurm, L. H. (2003). Language comprehension and production. In A. F. Healy & R. W. Proctor (Eds.), *Comprehensive Handbook of Psychology*, *Volume 4: Experimental Psychology*. New York: John Wiley & Sons.

Treutlein, A. (2011). *Rekodieren im Deutschen und Englischen: Wie rekodieren Englischlerner/-innen mit deutscher Muttersprache englische Wörter?* Dissertation, Eberhard Karls Universitat Tubingen.

Tversky, B. (1973). Encoding processes in recognition and recall. *Cognitive psychology*, *5*(3), 275–287.

Ulmi, M., Bürki, G. Verhein, A. & Marti, M. (2014). *Textdiagnose und Schreibberatung: Fach- und Qualifizierungsarbeiten begleiten*. Opladen, Toronto: Budrich.

Urban, D. & Mayerl, J. (2006). *Regressionsanalyse: Theorie, Technik und Anwendung.* 2. Auflage. Wiesbaden: VS Verlag für Sozialwissenschaften.

Valtin, R. (Hrsg.) (2002). *Was ist ein gutes Zeugnis?: Noten und verbale Beurteilungen auf dem Prüfstand*. Weinheim: Beltz Juventa.

van der Leeden, R. & Busing, F. M. (1994). *First iteration versus final IGLS/RIGLS estimators in two-level models: A Monte Carlo study with ML3*. Department of Psychology, University of Leiden.

van Dijk, T. A. (1980). *Textwissenschaft: eine interdisziplinäre Einführung.* Tübingen: Niemeyer.

Varnhagen, C. K. (2000). Shoot the messenger and disregard the message? Children's attitudes toward spelling. *Reading Psychology*, *21*(2), 115–128.

Vater, H. (1992). *Einführung in die Textlinguistik: Struktur, Thema und Referenz in Texten.* München: Fink.

Vater, H. (2001). *Einführung in die Textlinguistik* (3. Auflage). München: Fink.

Veal, L. R. & Tillman, M. (1971). Mode of discourse variation in the evaluation of children's writing. *Research in the Teaching of English*, *5*(1), 37–45.

Verhoeven, L. & Carlisle, J. (2006). Morphology in word identification and word spelling. *Reading and Writing, 19*, 643–650.

Verhoeven, L. & Perfetti, C. A. (2011). Morphological processing in reading acquisition: A cross-linguistic perspective. *Applied Psycholinguistics*, *32*(03), 457–466.

Verhoeven, L. & van Leeuwe, J. (2009). Modeling the growth of word decoding skills: Evidence from Dutch. *Scientific Studies of Reading, 13*, 205–223.

Viehweger, D. (1983). Textlinguistik. In W. Fleischer, Hartung, W., Schildt, J., Suchsland, P. (Hrsg.), *Deutsche Sprache: Kleine Enzyklopädie* (S. 211–237). Leipzig: VEB Bibliographisches Institut.

Völzing, P. L. (1980). Argumentation. Ein Forschungsbericht. *Zeitschrift für Literaturwissenschaft und Linguistik*, *10*, 204–235.

Walter, G. (2003). *Sprache – der Schlüssel zur Welt.* Freiburg im Breisgau: Herder.

Wang, W. C. (2004). Direct estimation of correlation as a measure of association strength using multidimensional item response models. *Educational and psychological measurement*, *64*(6), 937–955.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450.

Wasserman, J. D. & Bracken, B. A. (2003). Psychometric characteristics of assessment procedures. In I. Weiner, D. Freedheim, J. Schinka & W. Velicer (Eds.), *Handbook* of *psychology* (43–66). New York: Wiley.

Weber, A. (1973). *Dialektik der Aufsatzbeurteilung*. Donauwörth: Auer.

Weidenmann, B. (1997). "Multimedia": Mehrere Medien, mehrere Codes, mehrere Sinneskanäle? *Unterrichtswissenschaft*, *25*(3), 197–206.

Weinert, F. E. (Hrsg.) (2001). *Leistungsmessung in Schulen*. Weinheim, Basel: Beltz.

Weir, C. J. (1990). *Communicative language testing.* Hempel Hempstead, UK: Prentice Hall.

Weisberg, R. W. (1986). *Creativity: Genius and other myths.* New York: Freeman.

Weiss, R. (1965). Über die Zuverlässigkeit der Ziffernbenotung bei Aufsätzen. *Schule und Psychologie*, Heft 9, 257–269.

Weiss, R. (1966). Über die Auswirkung bestimmter Einstellungen auf Zensuren. *Unser Weg*, 166–177.

Welke, K. (1993). *Funktionale Satzperspektive. Ansätze und Probleme der funktionalen Grammatik.* Münster: Nodus Publikationen.

Westen, D. & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology*, *84*(3), 608–618.

White, E. M. (1984). Holisticism. *College Composition and Communication 35*(4), 400–409.

White, E. M. (1985). *Teaching and Assessing Writing: Recent Advances in Understanding, Evaluating, and Improving Student Performance.* San Francisco, CA: Jossey Bass.

White, L. (1989). The principle of adjacency in second language acquisition: do L2 learners observe the subset principle? In S. Gass & J. Schachter (Eds.), *Linguistic Perspectives on Second Language Acquisition* (pp. 134–158). Cambridge: Cambridge University Press.

White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychological Review, 66*, 297–333.

Wilcoxon, F. (1945). *Individual comparisons by ranking methods. Biometrics Bulletin 1*(6), 80–83.

Wilhelm, O. & Kunina, D. P. O. (2009). Pädagogisch-psychologische Diagnostik. In E. Wild & J. Möller, (Hrsg.), *Pädagogische Psychologie* (S. 307–331). Berlin, Heidelberg: Springer.

Wilkening, F. (2006). Informationsverarbeitungs-Theorien zur kognitiven Entwicklung. In W. Schneider & F. Wilkening (Hrsg.), *Theorien, Modelle und Methoden der Entwicklungspsychologie. Enzyklopädie der Psychologie* (Bd. C-V-1, S. 265–310). Göttingen: Hogrefe.

Willenberg, H. (2007). Lesen. In E. Klieme & B. Beck (Hrsg.), *Sprachliche Kompetenzen – Konzepte und Messung* (S. 107–117). Weinheim: Beltz.

Wimmer, G. & Altmann, G. (1996). The Theory of Word Length Distribution: Some Results and Generalizations. In P. Schmidt (Hrsg.), *Glottometrika 15* (S. 112–133). Trier: Wissenschaftlicher Verlag.

Wimmer, G., Köhler, R., Grotjahn, R. & Altmann, G. (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics 1*, 98–106.

Wimmer, H. (1993). Characteristics of developmental dyslexia in a regular writing system. *Applied Psycholinguistics, 14*, 1–33.

Winkelmann, H. & Böhme, K. (2009). Anlage und Durchführung der Pilotierung der Bildungsstandards. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik* (S. 31–41). Weinheim: Beltz.

Winkelmann, H. & Groeneveld, I. (2010). Geschlechterdisparitäten. In O. Köller, M. Knigge & B. Tesch (Hrsg.), *Sprachliche Kompetenzen im Ländervergleich* (S. 177–184). Münster: Waxmann.

Winther, E. (2010). *Kompetenzmessung in der beruflichen Bildung*. Bielefeld: Bertelsmann.

Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität.* Göttingen: Hogrefe.

Wise, S. L. (2009). Strategies for managing the problem of unmotivated examinees in low-stakes testing programs. *The Journal of General Education, 58*, 152–166.

Wolf, L. F. & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education, 8*, 227–242.

Wolfe, M. B. & Mienko, J. A. (2007). Learning and memory of factual content from narrative and expository text. *British Journal of Educational Psychology*, *77*(3), 541–564.

Wolfer, M. (2010). *Diagnostische Pädagogik als Grundlage für die (innere) Differenzierung zwischen Lernbehinderung und Hochbegabung*. Berlin: Logos.

„Wortschatz Universität Leipzig“ (2014). Letzter Zugriff am 30.03.2015 unter http://wortschatz.uni-leipzig.de/

Wu, M., Adams, R. J. & Wilson, M. R. (1998). *ConQuest: Generalized item response modeling software [Computer-programm]*. Camberwell, Victoria: Australian Council for Educational Research.

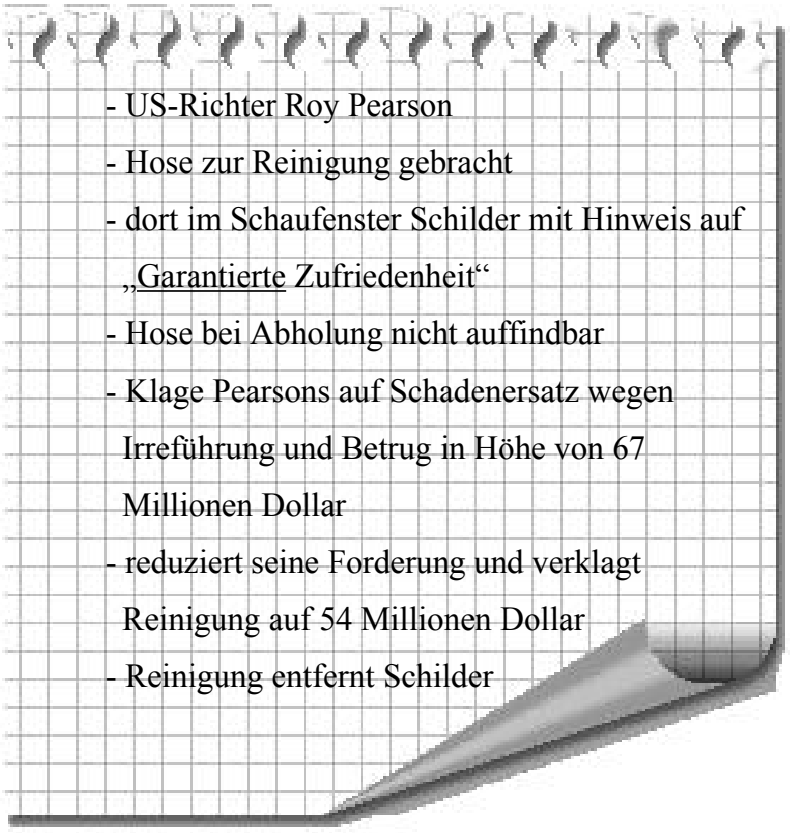
Yousfi, S. (2011). Methoden der Item- und Skalenkonstruktion. In: L. F. Hornke, M. Amelang, M. Kersting, N. Birbaumer, D. Frey, J. Kuhl, W. Schneider & R. Schwarzer (Hrsg.), *Methoden der psychologischen Diagnostik* (S. 151–211). Göttingen: Hogrefe.

Ziegler, J. C., Perry, C. & Coltheart, M. (2000). The DRC model of visual word recognition and reading aloud: An extension to German. *European Journal of Cognitive Psychology, 12*, 413–430.

Anhang

Anhang A.3.1.1: ***Aufgabenstimulus der informierenden Aufgabe „Zeitungsnachricht“.***

Der Reporter einer Tageszeitung soll für die morgige Ausgabe in der Rubrik „Aus aller Welt“ eine Nachricht schreiben. Dazu hat er sich folgende Stichpunkte notiert:



Schreibe die Nachricht für den Reporter und beachte dabei, dass in einem Artikel am Anfang die wichtigsten Informationen kommen und gegen Ende die unwichtigeren. Denk auch an die Überschrift!

Tipp:

Nummeriere zuerst die Informationen auf dem Notizzettel nach ihrer Wichtigkeit und schreibe dann.

Achte darauf, dass du mehrere Absätze schreibst.

Anhang A.3.1.2: ***Aufgabenstimulus der argumentierenden Aufgabe „Leserbrief“.***

In der Tageszeitung liest du folgenden Leserbrief.

Der Anlass zu diesem Brief ist eine Busfahrt, die ich heute gegen 13 Uhr mit der Linie B6 in Richtung Grombühl unternommen habe. Wieder einmal war ich empört über das ungebührliche und freche Verhalten von Jugendlichen, zumeist Schülern.

Schon beim Einsteigen in den Bus drängten sie mich unter Schreien und Toben zur Seite, so dass ich nur mit Mühe eine Fahrkarte lösen konnte. Als ich endlich ins Wageninnere gelangte, dachte natürlich keiner der jungen „Damen und Herren“ daran, mir einen Platz anzubieten, obwohl ich zwei schwere Einkaufstaschen zu schleppen hatte.

Die Jugendlichen unterhielten sich und lachten so laut, dass man sein eigenes Wort nicht verstehen konnte.

Als ich mir erlaubte zu sagen, dass es ihnen wohl an einer guten Kinderstube fehle, bekam ich nur freche Antworten und Gelächter zu hören.

Solches Verhalten hätte es zu meiner Jugendzeit nicht gegeben. Da trat man älteren Menschen noch in höflicher und rücksichtsvoller Weise gegenüber. Diese Verrohung der heutigen Jugend stimmt mich doch sehr bedenklich. Bringen denn die modernen Eltern und Lehrer den Jugendlichen überhaupt keine Manieren mehr bei?

Lina K.
97076 Würzburg
Veilchenweg 8

Schreibe einen Leserbrief, in dem du zu dem Vorwurf Lina K.s Stellung nimmst.

Du kannst dich vorbereiten, indem du drei Argumente für deine Position stichpunktartig formulierst!

1. _____

2. _____

3. _____

Schreibe nun den Brief für die Tageszeitung (nutze auch deine Stichpunkte). Achte darauf, dass du mehrere Absätze schreibst.

Anhang A.3.3.1: *Globalskala für argumentierende Texte.*

Stufe	Beschreibung
5	Der Schülertext ist in Bezug auf den Schreibanlass zu beurteilen. Ein Text dieser Kategorie weist folgende Merkmale auf: • Der Text bezieht einen klaren Standpunkt und entwickelt ihn mit stimmigen Begründungen und/oder Beispielen oder der Text entwickelt nachvollziehbar in einer abwägenden Argumentation eine vermittelnde Position. • Der Text verfügt über einen klaren Aufbau und weist kaum Mängel in der Kohärenz und/oder Unstimmigkeiten in der Argumentationsfolge auf. Die Textsorte wird durchweg eingehalten. • Der Text zeichnet sich durch abwechslungsreichen Satzbau und meist treffende Wortwahl aus. • Fehler in der Grammatik, in der Orthografie oder in der Kommasetzung treten kaum auf und beeinträchtigen nicht das Verständnis des Schülertextes.
4	Der Schülertext ist in Bezug auf den Schreibanlass zu beurteilen. Ein Text dieser Kategorie weist folgende Merkmale auf: • Der Text bezieht einen klaren Standpunkt und stützt ihn mit einigen Begründungen und/oder Beispielen oder der Text entwickelt in einer abwägenden Argumentation eine vermittelnde Position. • Der Text verfügt über einen klaren Aufbau, kann aber vereinzelt Unstimmigkeiten in der Argumentationsfolge, wenige Mängel in der Kohärenz und im Ausbau der Übergänge aufweisen. Die Textsorte wird größtenteils eingehalten. • Der Text demonstriert Sicherheit im Satzbau und in der Beachtung der Satzgrenzen, Satzbau und Wortwahl sind häufig einfach und wenig vielfältig. • Fehler in der Grammatik, in der Orthografie oder in der Kommasetzung treten vereinzelt auf, beeinträchtigen das Verständnis des Schülertextes jedoch kaum.
3	Der Schülertext ist in Bezug auf den Schreibanlass zu beurteilen. Ein Text dieser Kategorie weist eines oder mehrere der folgenden Merkmale auf: • Der Text bezieht einen Standpunkt und bemüht sich, ihn zu stützen, ist dabei aber nicht deutlich, sondern wiederholend, aufreihend oder argumentativ kaum entfaltet. • Der Text ist teilweise strukturiert, die Aussagen wirken gelegentlich unverbunden. Teilweise wird von der zutreffenden Textsorte abgewichen. • Der Text zeigt meist Sicherheit im Satzbau und in der Beachtung von Satzgrenzen, die Wortwahl ist an manchen Stellen unangemessen und/oder unzutreffend. • Fehler in der Grammatik, in der Orthografie oder in der Kommasetzung beeinträchtigen das Verständnis des Schülertextes an manchen Stellen.

2	Der Schülertext ist in Bezug auf den Schreibanlass zu beurteilen. Ein Text dieser Kategorie weist eines oder mehrere der folgenden Merkmale auf: • Der Text bezieht einen Standpunkt, aber die Stellungnahme wirkt sehr unverständlich, sie ist argumentativ kaum entfaltet oder stark wiederholend. • Dem Text mangelt es in erheblichem Maße an Struktur, die Aussagen sind dürftig miteinander verbunden oder die Erarbeitung ist zu knapp, um eine Struktur erkennen zu lassen. Von der zutreffenden Textsorte wird häufig abgewichen. • Der Text weist kaum Sicherheit im Satzbau und in der Beachtung von Satzgrenzen auf. Die Wortwahl ist oftmals unangemessen und/oder unzutreffend. • Fehler in der Grammatik oder im Sprachgebrauch, in der Orthografie und in der Kommasetzung beeinträchtigen das Verständnis des Schülertextes in vielen Teilen.
1	Der Schülertext ist in Bezug auf den Schreibanlass zu beurteilen. Ein Text dieser Kategorie weist eines oder mehrere der folgenden Merkmale auf: • Der Text bemüht sich, einen Standpunkt zu beziehen, doch die Erarbeitung ist unzusammenhängend, oder bezieht einen Standpunkt, doch begründet ihn nicht, oder paraphrasiert lediglich die Aufgabenstellung. • Der Text zeigt keine erkennbare Struktur oder besteht aus einer einzigen isolierten Aussage. Von der zutreffenden Textsorte wird abgewichen. • Der Text weist kaum oder überhaupt keine Beherrschung des Satzbaus und kaum oder gar keine Beachtung der Satzgrenzen auf. Die Wortwahl ist überwiegend oder über den gesamten Text hinweg unzutreffend. • Eine Vielfalt an Fehlern in Grammatik oder Sprachgebrauch, Orthografie und in der Kommasetzung verhindert in weiten Teilen das Verständnis des gesamten Schülertextes.
0	• Der Text ist zu kurz und bietet keine hinreichende Substanz für eine zuverlässige Bewertung.

Anhang A.3.3.2: *Globalskala für informierende Texte.*

Stufe	Beschreibung
5	Der Schülertext ist in Bezug auf den Schreibanlass zu beurteilen. Ein Text dieser Kategorie weist folgende Merkmale auf: • Der Text entwickelt und gestaltet die Informationen mit gut gewählten Details oder der Text stellt alle wichtigen gegebenen Informationen präzise dar. • Der Text verfügt über einen gelungenen Aufbau und ist durchgängig kohärent. Die Textsorte wird durchweg eingehalten. • Der Text zeichnet sich durch abwechslungsreichen Satzbau und meist treffende Wortwahl aus. • Fehler in der Grammatik, in der Orthografie oder in der Kommasetzung treten kaum auf und beeinträchtigen nicht das Verständnis des Schülertextes.
4	Der Schülertext ist in Bezug auf den Schreibanlass zu beurteilen. Ein Text dieser Kategorie weist folgende Merkmale auf: • Der Text entwickelt und gestaltet die Informationen teilweise mit Details oder der Text stellt nahezu alle wichtigen gegebenen Informationen dar. • Der Text verfügt über einen klaren Aufbau, kann aber einige wenige Mängel in der Kohärenz und/oder vereinzelt auftretende Unstimmigkeiten in der kausalen oder temporalen Abfolge aufweisen. Die Textsorte wird größtenteils eingehalten. • Der Text demonstriert Sicherheit im Satzbau und in der Beachtung der Satzgrenzen, Satzbau und Wortwahl sind aber eher einfach und wenig vielfältig. • Fehler in der Grammatik, in der Orthografie oder in der Kommasetzung treten vereinzelt auf, beeinträchtigen das Verständnis des Schülertextes jedoch kaum.
3	Der Schülertext ist in Bezug auf den Schreibanlass zu beurteilen. Ein Text dieser Kategorie weist folgende Merkmale auf: • Der Text stellt einige Informationen deutlich dar, aber wirkt auflistend, wenig entwickelt oder wiederholend oder der Text stellt nur einige der wichtigen gegebenen Informationen dar. • Der Text ist unregelmäßig strukturiert, die Aussagen wirken gelegentlich unverbunden. Teilweise wird von der zutreffenden Textsorte abgewichen. • Der Text zeigt meist Sicherheit im Satzbau und in der Beachtung von Satzgrenzen, die Wortwahl ist an manchen Stellen unangemessen oder unzutreffend. • Fehler in der Grammatik, in der Orthografie oder in der Kommasetzung beeinträchtigen das Verständnis des Schülertextes an manchen Stellen.

2	Der Schülertext ist in Bezug auf den Schreibanlass zu beurteilen. Ein Text dieser Kategorie weist eines oder mehrere der folgenden Merkmale auf: • Der Text stellt die Informationen nur fragmentarisch dar oder ist stark wiederholend oder kaum entwickelt. • Dem Text mangelt es in erheblichem Maße an Struktur, die Aussagen sind dürftig miteinander verbunden oder die Erarbeitung ist zu knapp, um eine Struktur erkennen zu lassen. Von der zutreffenden Textsorte wird häufig abgewichen. • Der Text weist kaum Sicherheit im Satzbau und in der Beachtung von Satzgrenzen auf. Die Wortwahl ist oftmals unangemessen oder unzutreffend. • Fehler in der Grammatik oder im Sprachgebrauch, in der Orthografie und in der Kommasetzung beeinträchtigen das Verständnis des Schülertextes in vielen Teilen.
1	Der Schülertext ist in Bezug auf den Schreibanlass zu beurteilen. Ein Text dieser Kategorie weist eines oder mehrere der folgenden Merkmale auf: • Der Text bemüht sich um eine Bearbeitung der Aufgabe, gibt dabei jedoch kaum oder keine zusammenhängenden Informationen oder paraphrasiert lediglich die Aufgabenstellung. • Der Text zeigt keine erkennbare Struktur oder besteht aus einer einzigen isolierten Aussage. Von der zutreffenden Textsorte wird abgewichen. • Der Text weist kaum oder überhaupt keine Beherrschung des Satzbaus und kaum oder gar keine Beachtung der Satzgrenzen auf. Die Wortwahl ist überwiegend oder über den gesamten Text hinweg unzutreffend. • Eine Vielfalt an Fehlern in Grammatik oder Sprachgebrauch, Orthografie und in der Kommasetzung verhindert in weiten Teilen das Verständnis des gesamten Schülertextes.
0	• Der Text ist zu kurz und bietet keine hinreichende Substanz für eine zuverlässige Bewertung.

Anhang A.3.3.3: *Globalskala für narrative Texte.*

Stufe	Beschreibung
5	Der Schülertext ist in Bezug auf den Schreibanlass zu beurteilen. Ein Text dieser Kategorie weist folgende Merkmale auf: • Der Text entfaltet eine gelungene Handlungsfolge, die durch gut gewählte Details entwickelt und gestaltet wird. • Der Text verfügt über einen stimmigen Aufbau mit klarer Erzählstruktur und gut ausgebauten Übergängen zwischen den Teilen der Handlung und ist durchgängig kohärent. Die Textsorte wird durchweg eingehalten. • Der Text zeichnet sich durch abwechslungsreichen Satzbau und meist treffende Wortwahl aus. • Fehler in der Grammatik, in der Orthografie oder in der Kommasetzung treten kaum auf und beeinträchtigen nicht das Verständnis des Schülertextes.
4	Der Schülertext ist in Bezug auf den Schreibanlass zu beurteilen. Ein Text dieser Kategorie weist folgende Merkmale auf: • Der Text entfaltet eine Handlungsfolge, die durch einige Details entwickelt und gestaltet wird. • Der Text verfügt über einen klaren Aufbau, kann aber vereinzelt Unstimmigkeiten in der Erzählfolge, wenige Mängel in der Kohärenz und im Ausbau der Übergänge aufweisen. Die Textsorte wird größtenteils eingehalten. • Der Text demonstriert Sicherheit im Satzbau und in der Beachtung der Satzgrenzen, Satzbau und Wortwahl sind aber eher einfach und wenig vielfältig. • Fehler in der Grammatik, in der Orthografie oder in der Kommasetzung treten vereinzelt auf, beeinträchtigen das Verständnis des Schülertextes jedoch kaum.
3	Der Schülertext ist in Bezug auf den Schreibanlass zu beurteilen. Ein Text dieser Kategorie weist folgende Merkmale auf: • Der Text bemüht sich um die Entwicklung einer Handlungsfolge, bleibt dabei in Teilen jedoch undeutlich, wenig entfaltet, mitunter aufreihend und wiederholend. • Der Text ist ansatzweise aufgebaut, zeigt aber deutliche Schwächen in der Erzählstruktur. Teile der Handlung stehen unverbunden im Text. Teilweise wird von der zutreffenden Textsorte abgewichen. • Der Text zeigt meist Sicherheit im Satzbau und in der Beachtung von Satzgrenzen, die Wortwahl ist an manchen Stellen unangemessen und/oder unzutreffend. • Fehler in der Grammatik, in der Orthografie oder in der Kommasetzung beeinträchtigen das Verständnis des Schülertextes an manchen Stellen.

2	Der Schülertext ist in Bezug auf den Schreibanlass zu beurteilen. Ein Text dieser Kategorie weist eines oder mehrere der folgenden Merkmale auf: • Der Text bemüht sich um die Erarbeitung einer Handlung, ist bei diesem Versuch aber fragmentarisch, stark wiederholend oder kaum entwickelt. • Dem Text mangelt es in erheblichem Maße an Struktur oder die Erarbeitung ist zu knapp, um eine Struktur erkennen zu lassen. Von der zutreffenden Textsorte wird häufig abgewichen. • Der Text weist kaum Sicherheit im Satzbau und in der Beachtung von Satzgrenzen auf. Die Wortwahl ist oftmals unangemessen und/oder unzutreffend. • Fehler in der Grammatik oder im Sprachgebrauch, in der Orthografie und in der Kommasetzung beeinträchtigen das Verständnis des Schülertextes in vielen Teilen.
1	Der Schülertext ist in Bezug auf den Schreibanlass zu beurteilen. Ein Text dieser Kategorie weist eines oder mehrere der folgenden Merkmale auf: • Der Text bemüht sich um eine Bearbeitung der Aufgabe, stellt jedoch keine zusammenhängenden Inhalte bereit, oder der Text paraphrasiert lediglich die Aufgabenstellung. • Der Text zeigt keine erkennbare Struktur oder besteht aus einer einzigen isolierten Aussage. Von der zutreffenden Textsorte wird abgewichen. • Der Text weist kaum oder überhaupt keine Beherrschung des Satzbaus und kaum oder gar keine Beachtung der Satzgrenzen auf. Die Wortwahl ist überwiegend oder über den gesamten Text hinweg unzutreffend. • Eine Vielfalt an Fehlern in Grammatik oder Sprachgebrauch, Orthografie und in der Kommasetzung verhindert in weiten Teilen das Verständnis des gesamten Schülertextes.
0	• Der Text ist zu kurz und bietet keine hinreichende Substanz für eine zuverlässige Bewertung.

Anhang A.3.3.4: *Textmusterspezifisches Gerüst der Stilskala für argumentierende Texte.*

Zur Bewertung des Stils ist auf folgende Kriterien zu achten:

- zutreffende Textsorte (…)
- einheitliche Perspektive und ggf. Adressierung (…)
- Kohärenzstiftung / Einsatz von Verknüpfungsmitteln
- Wortwahl
- Vielfalt und Komplexität der Satzkonstruktion

Aufgabenspezifische Textelemente:

- …

Stufe 4
Texte der Stufe 4 halten die zutreffende Textsorte ein. Die Perspektive ist einheitlich und zutreffend. Aufgrund funktional eingesetzter Verknüpfungsmittel wirken die Texte durchgängig kohärent. Die Wortwahl ist stets angemessen. Der Satzbau ist abwechslungsreich und nicht übermäßig komplex. Aufgabenspezifische Textelemente sind vorhanden.
Stufe 3
Texte der Stufe 3 halten die zutreffende Textsorte ein. Die Perspektive ist einheitlich und zutreffend. Verknüpfungsmittel sind größtenteils funktional. Inkohärenzen treten nur vereinzelt auf. Die Wortwahl ist zumeist angemessen. Der Satzbau ist größtenteils abwechslungsreich und nicht übermäßig komplex. Aufgabenspezifische Textelemente sind mehrheitlich vorhanden.
Stufe 2
Texte der Stufe 2 halten die zutreffende Textsorte größtenteils ein. Die Perspektive ist vorwiegend einheitlich und zutreffend. Verknüpfungsmittel finden sich kaum bzw. sind selten funktional. Oftmals sind Teile des Textes inkohärent. Die Wortwahl ist mitunter unangemessen. Die Sätze sind teilweise monoton oder übermäßig komplex konstruiert. Aufgabenspezifische Textelemente sind oftmals nicht vorhanden.
Stufe 1
Texte der Stufe 1 verfehlen meistens die zutreffende Textsorte. Die Perspektive ist bisweilen unklar oder unzutreffend. Verknüpfungsmittel werden nicht oder nicht funktional eingesetzt. Textteile wirken unzusammenhängend. Inkohärenzen treten gehäuft auf. Die Wortwahl ist kaum angemessen. Die Sätze sind überwiegend monoton konstruiert oder unverständlich. Aufgabenspezifische Textelemente fehlen.
Stufe 0
Texte der Stufe 0 sind zu kurz und bieten keine hinreichende Substanz für eine zuverlässige Bewertung.

Anhang A.3.3.5: *Textmusterspezifisches Gerüst der Stilskala für informierende Texte.*

Zur Bewertung des Stils ist auf folgende Kriterien zu achten:

- zutreffende Textsorte (…)
- einheitliche Perspektive und ggf. Adressierung (…)
- Kohärenzstiftung / Einsatz von Verknüpfungsmitteln
- Wortwahl
- Vielfalt und Komplexität der Satzkonstruktion
- Tempusgebrauch (…)

Aufgabenspezifische Textelemente:

- …

Stufe 4
Texte der Stufe 4 halten die zutreffende Textsorte ein. Die Perspektive ist einheitlich und zutreffend. Aufgrund funktional eingesetzter Verknüpfungsmittel wirken die Texte durchgängig kohärent. Die Wortwahl ist stets angemessen. Der Satzbau ist abwechslungsreich und nicht übermäßig komplex. Der Tempusgebrauch ist konsistent und angemessen. Aufgabenspezifische Textelemente sind vorhanden.
Stufe 3
Texte der Stufe 3 halten die zutreffende Textsorte ein. Die Perspektive ist einheitlich und zutreffend. Verknüpfungsmittel sind größtenteils funktional. Inkohärenzen treten nur vereinzelt auf. Die Wortwahl ist zumeist angemessen. Der Satzbau ist größtenteils abwechslungsreich und nicht übermäßig komplex. Der Tempusgebrauch ist überwiegend konsistent und angemessen. Aufgabenspezifische Textelemente sind mehrheitlich vorhanden.
Stufe 2
Texte der Stufe 2 halten die zutreffende Textsorte größtenteils ein. Die Perspektive ist vorwiegend einheitlich und zutreffend. Verknüpfungsmittel finden sich kaum bzw. sind selten funktional. Oftmals sind Teile des Textes inkohärent. Die Wortwahl ist mitunter unangemessen. Die Sätze sind teilweise monoton oder übermäßig komplex konstruiert. Der Tempusgebrauch ist gelegentlich inkonsistent oder unangemessen. Aufgabenspezifische Textelemente sind oftmals nicht vorhanden.
Stufe 1
Texte der Stufe 1 verfehlen meistens die zutreffende Textsorte. Die Perspektive ist bisweilen unklar oder unzutreffend. Verknüpfungsmittel werden nicht oder nicht funktional eingesetzt. Textteile wirken unzusammenhängend. Inkohärenzen treten gehäuft auf. Die Wortwahl ist kaum angemessen. Die Sätze sind überwiegend monoton konstruiert oder unverständlich. Der Tempusgebrauch ist oftmals inkonsistent oder unangemessen. Aufgabenspezifische Textelemente fehlen.
Stufe 0
Texte der Stufe 0 sind zu kurz und bieten keine hinreichende Substanz für eine zuverlässige Bewertung.

Anhang A.3.3.6: *Textmusterspezifisches Gerüst der Stilskala für narrative Texte.*

Zur Bewertung des Stils ist auf folgende Kriterien zu achten:

- zutreffende Textsorte (…)
- einheitliche Perspektive und ggf. Adressierung (…)
- Kohärenzstiftung / Einsatz von Verknüpfungsmitteln
- Wortwahl
- Vielfalt und Komplexität der Satzkonstruktion
- Tempusgebrauch (…)

Aufgabenspezifische Textelemente:

- Verwendung rhetorischer Stilmittel (Metaphern, Vergleiche, Ironie etc.)

Stufe 4
Texte der Stufe 4 halten die zutreffende Textsorte ein. Die Erzählperspektive ist einheitlich und zutreffend. Aufgrund funktional eingesetzter Verknüpfungsmittel wirken die Texte durchgängig kohärent. Die Wortwahl ist stets angemessen. Der Satzbau ist abwechslungsreich und nicht übermäßig komplex. Der Tempusgebrauch ist konsistent und angemessen. Aufgabenspezifische Textelemente sind vorhanden.
Stufe 3
Texte der Stufe 3 halten die zutreffende Textsorte ein. Die Erzählperspektive ist einheitlich und zutreffend. Verknüpfungsmittel sind größtenteils funktional. Inkohärenzen treten nur vereinzelt auf. Die Wortwahl ist zumeist angemessen. Der Satzbau ist größtenteils abwechslungsreich und nicht übermäßig komplex. Der Tempusgebrauch ist überwiegend konsistent und angemessen. Aufgabenspezifische Textelemente sind mehrheitlich vorhanden.
Stufe 2
Texte der Stufe 2 halten die zutreffende Textsorte größtenteils ein. Die Erzählperspektive ist vorwiegend einheitlich und zutreffend. Verknüpfungsmittel finden sich kaum bzw. sind selten funktional. Oftmals sind Teile des Textes inkohärent. Die Wortwahl ist mitunter unangemessen. Die Sätze sind teilweise monoton oder übermäßig komplex konstruiert. Der Tempusgebrauch ist gelegentlich inkonsistent oder unangemessen. Aufgabenspezifische Textelemente sind teilweise vorhanden.
Stufe 1
Texte der Stufe 1 verfehlen meistens die zutreffende Textsorte. Die Erzählperspektive ist bisweilen unklar oder unzutreffend. Verknüpfungsmittel werden nicht oder nicht funktional eingesetzt. Textteile wirken unzusammenhängend. Inkohärenzen treten gehäuft auf. Die Wortwahl ist kaum angemessen Die Sätze sind überwiegend monoton konstruiert oder unverständlich. Der Tempusgebrauch ist oftmals inkonsistent oder unangemessen. Aufgabenspezifische Textelemente fehlen.
Stufe 0
Texte der Stufe 0 sind zu kurz und bieten keine hinreichende Substanz für eine zuverlässige Bewertung.

Anhang A.3.3.7: *Aufgaben- und textmusterübergreifende Skala ‚sprachliche Richtigkeit'.*

Zur Bewertung der sprachlichen Richtigkeit sind folgende Merkmale zu berücksichtigen*:

- Richtige Rechtschreibung
- Richtige Zeichensetzung:
  a) Markierung von Satzgrenzen (Satzschlusszeichen, Großschreibung am Satzanfang)
  b) Kommasetzung
- Grammatikalische Korrektheit,
  unter anderem bezüglich:
  c) Satzbau
  d) Genus- und Kasusgebrauch
  e) Beugung von Wörtern

Stufe 4

Texte der Stufe 4 erfüllen die Anforderungen an die sprachliche Richtigkeit entsprechend der oben angeführten Merkmale. Es treten nahezu keine Fehler in Orthografie, Grammatik und Kommasetzung auf. Die Satzgrenzen sind durchgängig richtig markiert, alle Sätze sind vollständig.

Stufe 3

Texte der Stufe 3 erfüllen größtenteils die Anforderungen an die sprachliche Richtigkeit entsprechend der oben angeführten Merkmale. Es treten kaum Fehler in Orthografie, Grammatik und Kommasetzung auf. Der Lesefluss wird nicht beeinträchtigt. Sehr selten sind Satzgrenzen falsch oder nicht markiert oder Sätze unvollständig.

Stufe 2

Texte der Stufe 2 erfüllen teilweise die Anforderungen an die sprachliche Richtigkeit entsprechend der oben angeführten Merkmale. Es treten Fehler in Orthografie, Grammatik und Kommasetzung auf, die den Lesefluss beeinträchtigen. Zum Teil sind Satzgrenzen falsch oder nicht markiert oder Sätze unvollständig.

Stufe 1

Texte der Stufe 1 erfüllen kaum die Anforderungen an die sprachliche Richtigkeit entsprechend der oben angeführten Merkmale. Es treten gehäuft grammatikalische und orthografische Fehler auf, die den Lesefluss unterbrechen. Hierzu zählen mehrfach nicht beachtete Satzgrenzen oder unvollständige Sätze. Die Kommasetzung gelingt kaum.

Stufe 0

Texte der Stufe 0 sind zu kurz und bieten keine hinreichende Substanz für eine zuverlässige Bewertung.

** Die Auflistung erfolgt nicht hierarchisch.*

Anhang A.3.3.8: *Aufgabenspezifische Inhaltsskala für die informierende Aufgabe „Zeitungsnachricht“.*

Zur Bewertung des Inhalts ist auf folgende Kriterien zu achten:

- vollständige Berücksichtigung der gegebenen Informationen
- inhaltliche Stimmigkeit in der Darstellung der Informationen unter Berücksichtigung ihrer Priorität (Haupt- und Zusatzinformationen) sowie – wenn vorhanden – der Prägnanz und Adäquatheit der Überschrift
- kein Einbringen eigener Aspekte und Ideen, die über die vorgegebenen Informationen hinausgehen und den Kontext des Vorfalls verfälschen

Stufe 4
Texte der Stufe 4 berücksichtigen die gegebenen Informationen vollständig. Die Darstellung der Informationen ist inhaltlich konsistent und stringent und sie folgt der Priorität der gegebenen Informationen. Es werden keine weiterführenden eigenen Ideen eingebunden.
Stufe 3
Texte der Stufe 3 berücksichtigen die gegebenen Informationen vollständig. Die Darstellung der Informationen ist inhaltlich zumeist stimmig, folgt aber teilweise nicht der Priorität der gegebenen Informationen, sondern eher der Abfolge des Geschehens. Es werden keine weiterführenden eigenen Ideen eingebunden.
Stufe 2
Texte der Stufe 2 berücksichtigen die wichtigen gegebenen Informationen. Die Darstellung der Informationen ist inhaltlich größtenteils stimmig, folgt aber nicht der Priorität der gegebenen Informationen. Selten werden weiterführende eigenen Ideen eingebunden, die den Kontext des Vorfalls verfälschen.
Stufe 1
Texte der Stufe 1 berücksichtigen nur einige der gegebenen Informationen. Die Darstellung der Informationen ist inhaltlich nicht stimmig. Häufig werden weiterführende eigene Ideen eingebunden.
Stufe 0
Texte der Stufe 0 sind zu kurz und bieten keine hinreichende Substanz für eine zuverlässige Bewertung.

Anhang A.3.3.9: *Aufgabenspezifische Stilskala für die informierende Aufgabe „Zeitungsnachricht“.*

Zur Bewertung des Stils ist auf folgende Kriterien zu achten:

- zutreffende Textsorte (Zeitungsbericht)
- einheitliche Perspektive
- Kohärenzstiftung / Einsatz von Verknüpfungsmitteln
- Wortwahl
- Vielfalt und Komplexität der Satzkonstruktion
- Tempusgebrauch (Präteritum - einfache Vergangenheitsform)

Aufgabenspezifische Textelemente:

- Überschrift

Anspruchsvolle Elemente wie rhetorische Mittel und Ironie sind in der Bewertung wohlwollend zu berücksichtigen.

Stufe 4
Texte der Stufe 4 halten die zutreffende Textsorte ein. Die Perspektive ist einheitlich und zutreffend. Aufgrund funktional eingesetzter Verknüpfungsmittel wirken die Texte durchgängig kohärent. Die Wortwahl ist stets angemessen. Der Satzbau ist abwechslungsreich und nicht übermäßig komplex. Der Tempusgebrauch ist konsistent und angemessen. Aufgabenspezifische Textelemente sind vorhanden.
Stufe 3
Texte der Stufe 3 halten die zutreffende Textsorte ein. Die Perspektive ist einheitlich und zutreffend. Verknüpfungsmittel sind größtenteils funktional. Inkohärenzen treten nur vereinzelt auf. Die Wortwahl ist zumeist angemessen. Der Satzbau ist größtenteils abwechslungsreich und nicht übermäßig komplex. Der Tempusgebrauch ist überwiegend konsistent und angemessen. Aufgabenspezifische Textelemente sind mehrheitlich vorhanden.
Stufe 2
Texte der Stufe 2 halten die zutreffende Textsorte größtenteils ein. Die Perspektive ist vorwiegend einheitlich und zutreffend. Verknüpfungsmittel finden sich kaum bzw. sind selten funktional. Oftmals sind Teile des Textes inkohärent. Die Wortwahl ist mitunter unangemessen. Die Sätze sind teilweise monoton oder übermäßig komplex konstruiert. Der Tempusgebrauch ist gelegentlich inkonsistent oder unangemessen. Aufgabenspezifische Textelemente sind oftmals nicht vorhanden.
Stufe 1
Texte der Stufe 1 verfehlen meistens die zutreffende Textsorte. Die Perspektive ist bisweilen unklar oder unzutreffend. Verknüpfungsmittel werden nicht oder nicht funktional eingesetzt. Textteile wirken unzusammenhängend. Inkohärenzen treten gehäuft auf. Die Wortwahl ist kaum angemessen. Die Sätze sind überwiegend monoton konstruiert oder unverständlich. Der Tempusgebrauch ist oftmals inkonsistent oder unangemessen. Aufgabenspezifische Textelemente fehlen.
Stufe 0
Texte der Stufe 0 sind zu kurz und bieten keine hinreichende Substanz für eine zuverlässige Bewertung.

Anhang A.3.3.10: *Aufgabenspezifische Inhaltsskala für die argumentierende Aufgabe „Leserbrief“.*

Zur Bewertung des Inhalts ist auf folgende Kriterien zu achten:

- einleitende Bezugnahme auf den Leserbrief von Lina K.
- Einführung eines Vorwurfs der Lina K. (Vorwurf an die Jugendlichen allgemein oder konkret auf das Erlebnis bezogen oder Vorwurf an Eltern und Lehrer)
- Positionierung hinsichtlich eines Vorwurfs der Lina K. (alle drei Positionen sind gleichwertig zu bewerten)
- argumentative Stützung der eigenen Position:
  - **zustimmend**: Die Position der Lina K. wird bestätigt und durch das Anführen weiterer Argumente, Beispiele und Erklärungen unterstützt.
  - **ablehnend**: Die Position der Lina K. wird abgelehnt. Gegenargumente, Gegenbeispiele und Erklärungen werden angeführt.
  - **vermittelnd**: Es wird sowohl Verständnis für die Position der Lina K. aufgebracht, als auch das Verhalten der Jugendlichen erklärt.
- Überzeugungskraft und Angemessenheit der Argumente
- abrundendes Schlusselement

Stufe 4
Texte der Stufe 4 beziehen sich ausdrücklich auf einen Vorwurf der Lina K. Die eigene Position ist deutlich, sie wird mit zutreffenden Beispielen und Begründungen gestützt. Die Argumentation ist durchweg überzeugend. Einführung und Schluss bieten einen sinnvollen Themenein- und -ausstieg.
Stufe 3
Texte der Stufe 3 beziehen sich auf einen der Vorwürfe von Lina K. Die eigene Position ist größtenteils deutlich und wird mit überwiegend zutreffenden Beispielen und Begründungen gestützt. Die Argumentation ist angemessen und größtenteils überzeugend. Die Einführung in die Thematik ist hinreichend. Ein abrundendes Schlusselement ist größtenteils vorhanden.
Stufe 2
Texte der Stufe 2 nehmen zumindest implizit Bezug auf einen der Vorwürfe von Lina K. Die eigene Position ist nur im Ansatz entwickelt und zum Teil mit Beispielen und Begründungen gestützt. Die Argumentation wirkt teilweise angemessen. Die Einleitung ist zu knapp. Ein abrundendes Schlusselement ist teilweise vorhanden.
Stufe 1
Texte der Stufe 1 bemühen sich um eine Bearbeitung, verfehlen aber den Schreibanlass. Die Aufgabenstellung oder der Leserbrief der Lina K. wird bspw. paraphrasiert und es wird keine eigene Position eingenommen. Eine argumentative Auseinandersetzung findet nicht statt. Eine Einführung in die Thematik sowie ein abrundendes Schlusselement fehlen.
Stufe 0
Texte der Stufe 0 sind zu kurz und bieten keine hinreichende Substanz für eine zuverlässige Bewertung.

Anhang A.3.3.11: *Aufgabenspezifische Stilskala für die argumentierende Aufgabe „Leserbrief“.*

Zur Bewertung des Stils ist auf folgende Kriterien zu achten:

- zutreffende Textsorte (argumentativer Leserbrief)
- einheitliche Perspektive (inkl. zutreffende Adressierung)
- Kohärenzstiftung / Einsatz von Verknüpfungsmitteln
- Wortwahl
- Vielfalt und Komplexität der Satzkonstruktion

Aufgabenspezifische Textelemente:

- Abschiedsgrußformel und/oder Autornennung

Additum: Sprachliche Stilmittel (Attribute, rhetorische Fragen, sprachliche Bilder, Zitate) sind positiv zu berücksichtigen.

Stufe 4
Texte der Stufe 4 halten die zutreffende Textsorte ein. Die Perspektive ist einheitlich und zutreffend. Aufgrund funktional eingesetzter Verknüpfungsmittel wirken die Texte durchgängig kohärent. Die Wortwahl ist stets angemessen. Der Satzbau ist abwechslungsreich und nicht übermäßig komplex. Aufgabenspezifische Textelemente sind vorhanden.
Stufe 3
Texte der Stufe 3 halten die zutreffende Textsorte ein. Die Perspektive ist einheitlich und zutreffend. Verknüpfungsmittel sind größtenteils funktional. Inkohärenzen treten nur vereinzelt auf. Die Wortwahl ist zumeist angemessen. Der Satzbau ist größtenteils abwechslungsreich und nicht übermäßig komplex. Aufgabenspezifische Textelemente sind mehrheitlich vorhanden.
Stufe 2
Texte der Stufe 2 halten die zutreffende Textsorte größtenteils ein. Die Perspektive ist vorwiegend einheitlich und zutreffend. Verknüpfungsmittel finden sich kaum bzw. sind selten funktional. Oftmals sind Teile des Textes inkohärent. Die Wortwahl ist mitunter unangemessen. Die Sätze sind teilweise monoton oder übermäßig komplex konstruiert. Aufgabenspezifische Textelemente sind oftmals nicht vorhanden.
Stufe 1
Texte der Stufe 1 verfehlen meistens die zutreffende Textsorte. Die Perspektive ist bisweilen unklar oder unzutreffend. Verknüpfungsmittel werden nicht oder nicht funktional eingesetzt. Textteile wirken unzusammenhängend. Inkohärenzen treten gehäuft auf. Die Wortwahl ist kaum angemessen. Die Sätze sind überwiegend monoton konstruiert oder unverständlich. Aufgabenspezifische Textelemente fehlen.
Stufe 0
Texte der Stufe 0 sind zu kurz und bieten keine hinreichende Substanz für eine zuverlässige Bewertung.

Anhang A.3.8.1: *Kompetenzstufenbeschreibungen des KSM Schreiben für argumentierende Texte (IQB, 2014, S. 12–17).*

Kompetenzstufe I (bis 369 Punkte; Mindeststandards verfehlt)

Schülerinnen und Schüler auf Kompetenzstufe I erfassen die Aufgabenstellung nur in Ansätzen, bemühen sich aber in der Regel um eine Bearbeitung der Aufgabe. Ist in der Aufgabe die zu vertretende Position bzw. die zu stützende These vorgegeben und eine lineare Argumentation verlangt, wird die geforderte Anzahl der Argumente nicht erreicht. Wird keine Position vorgegeben und muss zunächst die argumentative Struktur einer Textvorlage erfasst werden, gelingt dies den Schülerinnen und Schülern noch nicht. Sind Argumente vorgegeben, etwa in Form von Listen, können sie zumindest ansatzweise reproduziert werden. Wenn keine Position vorgegeben ist und die Argumente selbst zu entwickeln sind, wird die verlangte Anzahl der Argumente in der Regel nicht erbracht. Auch das besondere Gewicht eines oder mehrerer Argumente wird nicht markiert. In der Regel wird keine Konklusion bzw. Position formuliert. Enthält der Text eine Konklusion oder Position, etwa in Form einer Handlungsempfehlung, erscheint diese als nicht hinreichend gestützt.

Schülerinnen und Schüler auf dieser Stufe schreiben in der Regel so kurze Texte, dass der argumentative Duktus nur ansatzweise erkennbar ist. Die Texte sind zumeist unzureichend strukturiert. Einleitung und Schluss fehlen oft; wenn vorhanden, sind sie oftmals vom Hauptteil nicht grafisch-formal getrennt. Die Schülerinnen und Schüler verfehlen häufig die intendierte Textsorte. Textsortenspezifische Elemente wie Leseransprache, Überschrift, Grußformel oder Autornennung finden sich nur selten. Die Leseransprache ist oftmals unzutreffend, der Adressatenbezug inkonsistent. Verknüpfungsmittel wie kausale und adversative Konjunktionen (z. B. *weil, aber*) oder Adverbien werden selten verwendet. Als Leser hat man erhebliche Schwierigkeiten, das Geschriebene als kohärent anzusehen.

Der Stil ist in der Regel sachlich; subjektiv-emotional motivierte Aussagen finden sich selten. Der Wortschatz, soweit er nicht der Aufgabe entnommen werden konnte, ist allerdings durchweg schmal; einzelne Wörter werden oft wiederholt. Der Satzbau ist mehrheitlich einförmig, syntaktische Muster werden häufig wiederholt.

Mehrheitlich häufen sich in den Texten grammatische und orthografische Fehler. Zuweilen werden Sätze formal (d. h. durch Satzschlusszeichen) gar nicht abgegrenzt. Oftmals wird der Satzbau nicht oder nur unzureichend beherrscht. Die Kommaschreibung weist erhebliche Mängel auf. Einzelne Wörter und Wendungen werden häufig inkorrekt gebraucht.

Kompetenzstufe II (370 bis 464 Punkte; Mindeststandard)

Schülerinnen und Schüler auf Kompetenzstufe II können die Aufgabenstellung grundsätzlich erfassen und in Ansätzen argumentativ bearbeiten. Wird in der Aufgabe die Position bzw. die zu stützende These vorgegeben und eine lineare Argumentation verlangt, sind sie in der Lage, einzelne Argumente zu formulieren, die geforderte Mindestanzahl der Argumente wird in der Regel jedoch nicht erreicht. Wenn keine Position vorgegeben und zunächst die argumentative Struktur einer Textvorlage erfasst werden muss, gelingt das nur in Ansätzen. Sind Argumente, etwa in Form von Listen, vorgegeben, können die Schülerinnen und Schüler diese häufig reproduzieren. Sind die Argumente selbst zu entwickeln, wird die verlangte Anzahl der Argumente

zumindest manchmal erbracht. Das besondere Gewicht eines oder mehrerer Argumente wird jedoch fast durchgängig nicht markiert. Positionen bzw. Konklusionen werden bisweilen formuliert, sie folgen aber nicht schlüssig aus den übrigen Argumenten.

Auf dieser Stufe schreiben Schülerinnen und Schüler eher kürzere Texte, wobei aber der argumentative Duktus in Ansätzen erkennbar ist. Die Schülerinnen und Schüler sind jedoch noch nicht in der Lage, ihre Texte angemessen zu strukturieren. Häufig fehlen etwa Einleitung und/oder Schluss; wenn vorhanden, sind sie oftmals vom Hauptteil nicht grafisch-formal getrennt. Die intendierte Textsorte ist zumeist erkennbar. So wird in der Regel deutlich, dass es sich beispielsweise um einen förmlichen oder vertraulichen Brief handelt. Die Leseransprache ist mehrheitlich treffend, der Adressatenbezug weitgehend konsistent. Textsortenspezifische Elemente wie Leseransprache, Überschrift, Grußformel oder Autornennung enthalten die Texte aber nur selten. Verknüpfungsmittel wie kausale und adversative Konjunktionen (z. B. *weil, aber*) oder Adverbien werden gelegentlich verwendet. Als Leser hat man insgesamt jedoch häufig Schwierigkeiten, den Text bzw. die Textteile als kohärent anzusehen.

Der Stil ist weitgehend unangemessen. Der Wortschatz ist, soweit er nicht der Aufgabe entnommen werden konnte, in der Regel eher schmal, häufig werden einzelne Wörter wiederholt. Zudem ist der Satzbau nur selten variabel, syntaktische Muster werden häufig wiederholt.

In den Schülertexten häufen sich bisweilen grammatische und orthografische Fehler. Gelegentlich werden die Sätze formal (d. h. durch Satzschlusszeichen) nicht abgegrenzt. Der Satzbau wird oftmals nicht durchgängig beherrscht, die Kommaschreibung gelingt zuweilen auch in einfacheren Fällen nicht. Einzelne Wörter und Wendungen werden teilweise inkorrekt gebraucht.

Kompetenzstufe III (465 bis 559 Punkte; Regelstandard)

Auf Kompetenzstufe III bearbeiten die Schülerinnen und Schüler die Aufgabenstellung insofern zufriedenstellend, als dass in ihren Texten ein Standpunkt zumeist argumentativ plausibel gestützt wird. Ist in der Aufgabe die Position bzw. die zu stützende These vorgegeben und wird eine lineare Argumentation verlangt, werden zwar Argumente genannt, die geforderte Anzahl der Argumente wird aber nicht immer erreicht. Ansatzweise sind die Schülerinnen und Schüler in der Lage, ihre Argumentation auch durch weitergehende Ausführungen (Belege oder Beispiele) zu stützen. Wird keine Position vorgegeben und muss die argumentative Struktur einer Textvorlage zunächst erfasst werden, gelingt es den Schülerinnen und Schülern häufig, sich auf den inhaltlichen Kern zu beziehen und einzelne Argumente zu entwickeln. Oft können vorgebrachte Argumente mit Bezug auf eigene Erfahrungen, Beispiele, Belege, Zusatzinformationen, Vorschläge für Lösungsmöglichkeiten usw. gestützt werden. Sind Argumente vorgegeben, kann in etwa die Hälfte der Schülerinnen und Schüler sie im Hinblick auf das eigene Schreibziel nachvollziehbar arrangieren. Sind Argumente selbst zu entwickeln, wird die verlangte Anzahl an Argumenten in der Regel erbracht. Häufig werden Konklusionen oder Handlungsempfehlungen formuliert, wobei jedoch das besondere Gewicht eines oder mehrerer Argumente nur sehr selten markiert wird. Dennoch erscheint die resultierende Position als weitgehend gestützt.

Auf dieser Stufe sind Texte in der Regel hinreichend umfangreich, sodass der argumentative Duktus durchweg erkennbar ist. Die Schülerinnen und Schüler strukturieren ihre Texte in Ansätzen. So enthalten sie häufig

zumindest eine Einleitung und/oder ein Schluss, die zudem meist vom Hauptteil grafisch-formal getrennt sind. Nahezu alle Schülerinnen und Schüler formulieren textsortengemäß, etwa wenn sie einen förmlichen oder vertraulichen Brief verfassen sollen. Die Leseransprache ist in der Regel treffend, der Adressatenbezug zumeist konsistent. Textspezifische Elemente wie Überschrift, Grußformel oder Autornennung werden bisweilen formuliert. Verknüpfungsmittel wie kausale und adversative Konjunktionen (z. B. *weil, aber*) oder Adverbien werden jedoch nicht immer angemessen verwendet. Als Leser hat man in einigen Fällen Schwierigkeiten, Teile des Geschriebenen als kohärent anzusehen.

Auf dieser Stufe schreiben die Schülerinnen und Schüler bisweilen auf angemessenem stilistischem Niveau und hinreichend präzise. Der Wortschatz, soweit er nicht der Aufgabe entnommen werden konnte, ist manchmal recht umfangreich; Wiederholungen sind selten. In etwa der Hälfte der Texte ist der Satzbau variabel.

Schülerinnen und Schüler auf dieser Stufe zeigen weitestgehend Sicherheit im Satzbau und in der Beachtung von Satzgrenzen. Fehler in der Grammatik, in der Orthografie und in der Kommaschreibung treten bisweilen auf, beeinträchtigen das Verständnis jedoch in der Regel nicht. Mehrheitlich werden Wörter und Wendungen korrekt gebraucht.

Kompetenzstufe IV (560 bis 654 Punkte; Regelstandard plus)

Schülerinnen und Schüler auf Kompetenzstufe IV können ihre Texte so gestalten, dass der argumentative Zusammenhang in der Regel deutlich ist. Es wird dabei durchweg erkennbar Position bezogen, z. B. eine Handlungsempfehlung gegeben. Diese wird weitgehend plausibel gestützt. Ist in der Aufgabe die Position bzw. die zu stützende These vorgegeben und wird eine lineare Argumentation verlangt, sind die Schülerinnen und Schüler nahezu durchgängig in der Lage, die jeweils geforderte Mindestanzahl der Argumente zu präsentieren und die Argumentation in der Regel durch weitergehende Ausführungen (Beispiele usw.) anzureichern. Wenn keine Position vorgegeben ist und zunächst die argumentative Struktur einer Textvorlage erfasst werden muss, gelingt es den Schülerinnen und Schülern größtenteils, sich auf den inhaltlichen Kern zu beziehen und selbstständig mehrere Argumente zu entwickeln. Dabei rekurrieren sie auf eigene Erfahrungen, Informationen aus anderen Quellen, allgemeine und als konsensfähig angesehene Aussagen usw. Sind Argumente vorgegeben, etwa in Form von Listen, können sie durchweg nachvollziehbar arrangiert werden. Wenn keine Position vorgegeben ist und Argumente selbst zu entwickeln sind, wird weitgehend die in der Aufgabe verlangte Anzahl formuliert. Eine explizite Gewichtung der Argumente erfolgt jedoch nur in seltenen Fällen.

Der argumentative Duktus der mehrheitlich längeren Texte ist deutlich erkennbar; sie sind zudem für den Leser gut nachvollziehbar strukturiert. So gibt es oft eine Einleitung und einen Schluss, die vom Hauptteil meist grafisch-formal getrennt sind. Die Schülerinnen und Schüler schreiben textsortenkonform zum Beispiel einen förmlichen oder vertraulichen Brief. Dabei ist nicht nur die Leseransprache weitestgehend treffend, häufig finden sich auch textsortenspezifische Elemente wie Überschrift, Grußformel oder Autornennung. Verknüpfungsmittel werden in der Regel angemessen verwendet und als Leser hat man kaum Schwierigkeiten, das Geschriebene als kohärent anzusehen.

Der Stil ist mehrheitlich angemessen; die Satzstrukturen sind in der Regel variabel. Es kommen u. a. Satzgefüge verschiedener Typen vor. Der nicht dem Stimulus entnommene Wortschatz ist in den meisten Fällen breit und treffend. Nichtfunktionale syntaktische und lexikalische Wiederholungen kommen nur selten vor.

Schülerinnen und Schülern auf dieser Stufe unterlaufen kaum Fehler in Orthografie und Grammatik, sie zeigen Sicherheit im Satzbau und in der Beachtung von Satzgrenzen. Auch die Kommaschreibung, insbesondere bei Aufzählungen und zur Abtrennung von Nebensätzen, wird in der Regel gemeistert. Wörter und Wendungen werden weitgehend korrekt gebraucht.

Kompetenzstufe V (ab 655 Punkten; Optimalstandard)

Schülerinnen und Schüler auf der höchsten Kompetenzstufe können ihre Texte so gestalten, dass der argumentative Zusammenhang deutlich ist. Sie formulieren eine nachvollziehbar, argumentativ hinreichend gestützte, in der Sache abwägende, ggf. vermittelnde Position, die durch mehrere plausible Argumente gestützt wird. Die Anzahl der Argumente entspricht den aufgabenspezifischen Forderungen oder übersteigt diese. Die Argumentation ist nahezu immer durch weitergehende Ausführungen (Beispiele usw.) angereichert. Wenn die argumentative Struktur einer Textvorlage erfasst werden muss, gelingt es den Schülerinnen und Schülern, sich auf den inhaltlichen Kern zu beziehen und selbstständig mehrere plausible Argumente zu entwickeln. Dabei rekurrieren sie auf eigene Erfahrungen, Informationen aus anderen Quellen, allgemeine, als konsensfähig angesehene Aussagen usw. Vorgegebene Argumente können durchweg nachvollziehbar und zielführend arrangiert werden. Eine explizite Gewichtung der Argumente erfolgt jedoch auch auf Kompetenzstufe V nur in einigen Fällen.* Schlussfolgerungen (z. B. in Form von Handlungsempfehlungen) sind nahezu durchgängig vorhanden.

Der argumentative Duktus der mehrheitlich längeren Texte ist deutlich erkennbar; die Textsorte wird durchweg eingehalten. Die Texte sind für den Leser gut nachvollziehbar strukturiert, Einleitung und Schluss sind weitestgehend in hinreichendem Umfang vorhanden und wurden meist vom Hauptteil grafisch-formal getrennt. Die Leseransprache ist fast immer treffend. Textsortenspezifische Elemente wie Überschrift, Grußformel oder Autornennung werden weitgehend angeführt. Verknüpfungsmittel werden durchgängig angemessen verwendet. Als Leser hat man keine Schwierigkeiten, das Geschriebene als kohärent anzusehen.

Der Stil ist größtenteils angemessen. Der nicht dem Stimulus entnommene Wortschatz ist breit und treffend, die Satzstrukturen weitestgehend variabel. Nichtfunktionale syntaktische und lexikalische Wiederholungen kommen kaum vor.

Schülerinnen und Schüler auf Kompetenzstufe V zeigen Sicherheit im Satzbau und in der Beachtung von Satzgrenzen. Orthografische und grammatische Fehler treten nur vereinzelt auf. Die Kommaschreibung gelingt häufig auch in schwierigeren Fällen. Wörter und Wendungen werden weitestgehend korrekt gebraucht.

---

* Die explizite Markierung des Gewichts eines Arguments stellt ein inhaltliches Qualitätsmerkmal dar, welches jedoch nicht alternativlos ist. So stehen den Schülerinnen und Schülern auch diverse Prinzipien der Gewichtung durch Struktur (Abfolge der Argumente) zur Verfügung, wodurch auch eine implizite (nicht verbalisierte) Gewichtung möglich ist.

Anhang A.3.8.2: *Kompetenzstufenbeschreibungen des KSM Schreiben für informierende Texte (IQB, 2014, S. 18–22).*

Kompetenzstufe I (bis 382 Punkte; Mindeststandards verfehlt)

Schülerinnen und Schüler auf Kompetenzstufe I bemühen sich in der Regel um eine Bearbeitung der Aufgaben und erfassen die Aufgabenstellung in Ansätzen. Sie geben dabei jedoch nur wenige oder keine zusammenhängenden Informationen bzw. paraphrasieren lediglich die Aufgabenstellung. Sind die inhaltlichen Aspekte zu rekonstruieren oder müssen bildliche in verbale Informationen übersetzt werden, gelingt dies größtenteils nicht. Selbst wenn die Aufgabenstellung die zentralen Inhalte explizit vorgibt, werden diese nur teilweise aufgegriffen. Die wenigen thematisierten Inhalte werden in der Regel detailarm und in ihrer Abfolge unstrukturiert wiedergegeben. Oftmals werden auch nichtfunktionale Informationen hinzugefügt. Daher ist es für den Leser im Grunde nicht möglich, die intendierten komplexen Zusammenhänge nachzuvollziehen.

Die Schülerinnen und Schüler auf dieser Stufe schreiben in der Regel kurze Texte, verfehlen aber häufig die Textsorte und strukturieren die Texte meist unzureichend. Der Stil der Beschreibungen ist überwiegend sachlich, das Tempus ist weitgehend angemessen. Bei den Berichten kommen aber häufig nicht-funktionale narrative und emotionale Markierungen vor; zudem wird das Tempus (Präteritum bzw. Plusquamperfekt) nicht immer konsistent verwendet. Oftmals nehmen die Schreibenden eine unangemessene Perspektive ein oder wechseln diese nichtfunktional (beispielsweise von „du“ zu „ihr“). Textspezifische Elemente (wie etwa Überschrift, Grußformel, Autornennung) sind nur selten vorhanden. Werden Verknüpfungsmittel wie Konjunktionen, Adverbien usw. verwendet, sind diese oft nicht richtig gebraucht. Als Leser hat man insofern erhebliche Schwierigkeiten, das Geschriebene als kohärent anzusehen. Eine formale Gliederung in Absätze kommt kaum vor.

Der Wortschatz, soweit er nicht der Aufgabenstellung entnommen werden konnte, ist schmal; Wörter (insbesondere solche mit verknüpfender Funktion) werden häufig wiederholt. Der Satzbau ist größtenteils wenig variabel und monoton.

In der Mehrheit der Texte häufen sich grammatische und orthografische Fehler. Zuweilen werden Sätze formal, also durch Satzschlusszeichen, gar nicht abgegrenzt. Oftmals wird der Satzbau nicht oder nur unzureichend beherrscht. Die Kommaschreibung weist erhebliche Mängel auf. Einzelne Wörter und Wendungen werden häufig inkorrekt gebraucht.

Kompetenzstufe II (383 bis 473 Punkte; Mindeststandard)

Schülerinnen und Schüler auf Kompetenzstufe II greifen einige der zentralen Informationen aus der Text- bzw. Bildvorlage auf, es bleiben jedoch Lücken bestehen, die ein umfassendes Verständnis des Lesers verhindern. Sind die inhaltlichen Aspekte zu rekonstruieren oder müssen Bild- in Textinformationen übersetzt werden, gelingt dies nur für wenige dieser Inhalte. Gibt die Aufgabenstellung die zentralen Inhalte jedoch explizit vor, wird die Mehrheit dieser Inhalte im Schülertext aufgegriffen. Einige der Informationen werden bisweilen detailliert wiedergegeben, es mangelt jedoch an Genauigkeit, oftmals finden sich auch fehlerhafte Elemente. Ein sinnvolles Arrangement der Inhalte ist in etwa der Hälfte der Texte zu erkennen. Ist es jedoch erforderlich, eine

komplexe Ereignisabfolge berichtend zu rekonstruieren, werden die Informationen nahezu immer in einer wenig sinnvollen und nicht zweckdienlichen Abfolge thematisiert.

Schülerinnen und Schüler auf dieser Stufe produzieren mehrheitlich kurze Texte und schreiben wenig strukturiert. Der Textsorte wird allerdings weitgehend entsprochen. Die beschreibend-instruierenden Texte sind in angemessener Weise sachlich gehalten. Auch bei den Berichten überwiegt der sachliche Duktus, sie enthalten aber oftmals auch nicht-funktionale narrative und emotive Komponenten. Die Perspektive ist weitgehend angemessen, häufig jedoch nicht einheitlich. So finden sich nicht selten nichtfunktionale Perspektivwechsel (beispielsweise von „du“ zu „ihr“). Bei den Beschreibungen ist die Tempuswahl weitestgehend korrekt (Präsens), bei den Berichten werden bisweilen aber unangemessene Tempusformen gebraucht. Textspezifische Elemente (wie etwa Überschrift, Grußformel, Autornennung) sind gelegentlich vorhanden. Verknüpfungsmittel werden teilweise verwendet, die Texte weisen jedoch oftmals Brüche auf. Eine formale Strukturierung in Absätze kommt selten vor.

Die Wortwahl ist teilweise unangemessen oder unzutreffend und es kommen häufig Wortwiederholungen ohne erkennbare stilistische Funktion vor. Der Satzbau ist überwiegend einfach und monoton.

In den Schülertexten häufen sich bisweilen grammatische und orthografische Fehler. Gelegentlich werden die Sätze formal, also durch Satzschlusszeichen, nicht abgegrenzt. Der Satzbau wird oftmals nicht durchgängig beherrscht, die Kommaschreibung gelingt zuweilen auch in einfacheren Fällen nicht. Einzelne Wörter und Wendungen werden teilweise inkorrekt gebraucht.

Kompetenzstufe III (474 bis 564 Punkte; Regelstandard)

Schülerinnen und Schüler auf Kompetenzstufe III greifen in ihren Texten die Mehrzahl der zentralen Informationen aus der Text- bzw. Bildvorlage auf. Sind die inhaltlichen Aspekte zu rekonstruieren oder ist Bild- in Textinformation zu überführen, gelingt dies für viele der Inhalte. Einige Inhalte werden jedoch nur angerissen. Gibt die Aufgabenstellung die zentralen Inhalte explizit vor, werden die meisten dieser Inhalte im Schülertext aufgegriffen. Oftmals werden die Informationen detailliert wiedergegeben. Die Inhalte werden mehrheitlich sinnvoll arrangiert. Lediglich wenn eine komplexe Ereignisabfolge berichtend rekonstruiert werden muss, werden die Informationen größtenteils in einer wenig sinnvollen und nicht zweckdienlichen Abfolge thematisiert.

Der Textsorte wird weitestgehend entsprochen; manchmal – vor allem in den berichtenden Texten – werden aber nichtkonforme Bausteine (z. B. narrative oder kommentierende Elemente) verwendet. Die Perspektive ist größtenteils angemessen und einheitlich, Perspektivwechsel sind meist funktional (beispielsweise von „du“ zu „man“), nichtfunktionale Perspektivwechsel (beispielsweise von „du“ zu „ihr“) finden sich selten. Der Tempusgebrauch ist bei den beschreibenden Texten fast durchgängig, bei den Berichten weitgehend korrekt. Kommen andere für die Kohärenz der Texte wichtige Mittel (z. B. Konjunktionen oder Pronomen) vor, sind sie überwiegend richtig gebraucht. Als Leser kann man fast durchgängig Kohärenz herstellen. Textsortenspezifische Elemente (wie etwa Überschrift, Grußformel, Autornennung) werden gelegentlich angefügt. Die Texte werden mehrheitlich grafisch-formal gegliedert.

Die Wortwahl ist oft angemessen und hinreichend präzise, gelegentliche nichtfunktionale Wiederholungen wirken kaum störend. Der Satzbau ist häufig variabel, parataktische Passagen sind meist funktional.

Schülerinnen und Schüler dieser Stufe zeigen weitestgehend Sicherheit im Satzbau und in der Beachtung von Satzgrenzen. Fehler in der Grammatik, in der Orthografie und in der Kommaschreibung treten bisweilen auf, beeinträchtigen das Verständnis jedoch in der Regel nicht. Mehrheitlich werden Wörter und Wendungen korrekt gebraucht.

Kompetenzstufe IV (565 bis 655 Punkte, Regelstandard plus)

Schülerinnen und Schüler auf Kompetenzstufe IV greifen den Großteil der zentralen Informationen aus der Text- bzw. Bildvorlage auf. Sind die inhaltlichen Aspekte zu rekonstruieren oder ist Bild- in Textinformation zu übersetzen, gelingt dies für die meisten Inhalte. Gibt die Aufgabenstellung die zentralen Inhalte explizit vor, werden nahezu alle Informationen im Schülertext aufgegriffen. Viele der Inhalte werden detailliert wiedergegeben, einige wenige jedoch nur angerissen. Die Informationen werden größtenteils sinnvoll arrangiert. Lediglich wenn eine komplexe Ereignisabfolge berichtend rekonstruiert werden muss, werden die Inhalte oftmals in einer wenig zweckdienlichen Abfolge thematisiert.

Der Textsorte wird nahezu immer entsprochen, sehr selten und nur in geringem Umfang werden nichtkonforme Bausteine (bspw. narrative Elemente) verwendet. Die Perspektive ist fast durchgängig angemessen und einheitlich. Perspektivwechsel sind meist funktional (beispielsweise von „du“ zu „man“), nichtfunktionale Perspektivwechsel (beispielsweise von „du“ zu „ihr“) finden sich nur noch vereinzelt. Textsortenspezifische Elemente wie Überschrift oder Autornennung werden in etwa der Hälfte der Texte angefügt. Der Tempusgebrauch ist weitestgehend konsistent und angemessen. Die meisten Texte sind kohärent, Verknüpfungsmittel werden größtenteils richtig gebraucht, die große Mehrheit der Texte ist grafisch-formal sinnvoll gegliedert.

Der Wortschatz ist weitestgehend umfangreich, die Wortwahl größtenteils angemessen und hinreichend präzise. Es finden sich nur wenige nichtfunktionale Wiederholungen. Die Sätze sind in der Regel variabel konstruiert.

Schülerinnen und Schülern auf dieser Stufe unterlaufen kaum Fehler in Orthografie und Grammatik, sie zeigen Sicherheit im Satzbau und in der Beachtung von Satzgrenzen. Auch die Kommaschreibung, insbesondere bei Aufzählungen und zur Abtrennung von Nebensätzen, wird in der Regel gemeistert. Wörter und Wendungen werden weitgehend korrekt gebraucht.

Kompetenzstufe V (ab 656 Punkten; Optimalstandard)

Schülerinnen und Schüler auf Kompetenzstufe V greifen alle zentralen Informationen aus der Text- bzw. Bildvorlage auf. Die meisten Inhalte werden angemessen detailliert wiedergegeben, die Details sind größtenteils korrekt und zweckdienlich. Die Inhalte sind weitgehend zutreffend und zielführend arrangiert. Selten treten kleinere Ungenauigkeiten in der Detaillierung oder im Arrangement auf. Lediglich wenn eine komplexe Ereignisabfolge berichtend rekonstruiert werden muss, finden sich bisweilen eingeschränkt zweckdienliche Informationsabfolgen.

Der Textsorte wird durchweg entsprochen. Die Perspektive ist durchgängig zutreffend und einheitlich; Perspektivwechsel, wenn vorhanden, sind funktional (beispielsweise von „du“ zu „man“). Der Tempusgebrauch ist fast durchgängig konsistent und angemessen. Die Texte sind kohärent, Verknüpfungsmittel werden richtig

gebraucht, die Gliederung ist weitestgehend grafisch verdeutlicht. Textspezifische Elemente (wie etwa Überschrift, Grußformel, Autornennung) kommen mehrheitlich vor.

Der Wortschatz ist umfangreich, teilweise sehr elaboriert. Die Wortwahl ist fast durchgängig angemessen und hinreichend präzise, nichtfunktionale Wiederholungen sind sehr selten. Der Satzbau ist mehrheitlich sehr abwechslungsreich.

Schülerinnen und Schüler auf Kompetenzstufe V zeigen Sicherheit im Satzbau und in der Beachtung von Satzgrenzen. Orthografische und grammatische Fehler treten nur vereinzelt auf. Die Kommaschreibung gelingt häufig auch in schwierigeren Fällen. Wörter und Wendungen werden weitestgehend korrekt gebraucht.

Anhang A.3.8.3: *Kompetenzstufenbeschreibungen des KSM Schreiben für narrative Texte (IQB, 2014, S. 24–29).*

Kompetenzstufe I (bis 411 Punkte; Mindeststandards verfehlt)

Schülerinnen und Schüler auf Kompetenzstufe I stellen zumeist einen Bezug zu den Aufgaben her, werden den Anforderungen aber nicht oder nur ansatzweise gerecht. In manchen Fällen wird die Aufgabenstellung falsch verstanden und es wird gar nicht erzählt. Ist eine prototypische „Höhepunkt"-Geschichte gefragt, sind konstitutive Elemente wie Komplikation und Auflösung nicht oder nur rudimentär erkennbar. Auch fehlt es häufig an einer Exposition (Einführung der Handelnden und ggf. des Schauplatzes und der Zeit) und einer Coda (erzähltypische Abschlusspassage). Enthält die Aufgabenstellung diverse inhaltliche Vorgaben, Hintergrundinformationen oder Mustervorlagen, orientieren sich die Schreibenden in der Regel an keinem oder nur an einem der dadurch vorgegebenen inhaltlichen Aspekte und greifen diesen im eigenen Text auf. Oftmals wird lediglich paraphrasiert. Die Schülerinnen und Schüler schreiben detailarm und gehen nur vereinzelt auf Gedanken, Gefühle und Handlungsmotive ein. Erzählrelevante beschreibende Anteile werden nicht oder nur rudimentär ausgearbeitet und sind häufig nicht zielführend umgesetzt.

Wenn ein narrativer Duktus erkennbar ist, gelingt es den Schülerinnen und Schülern gelegentlich, eine Erste- bzw. Dritte-Person-Perspektive konsistent durchzuhalten. Die Linearisierung von Ereignissen zu kohärenten Ereignisfolgen wird dagegen weitestgehend nicht gemeistert, es resultieren semantische Lücken. Darüber hinaus werden Teilaufgaben, die eine inhaltliche und sprachliche Unterscheidung von Exposition, Komplikation und Auflösung erfordern, größtenteils nicht bewältigt. Daher gelingt die Orientierung des Lesers im Hinblick auf Aktanten, Ort und Zeit überwiegend nicht. Eine Komplikation wird zumeist nicht markiert. Die in der Regel kurzen Texte wirken oftmals nicht abgeschlossen. Eine äußere und inhaltlich plausible Gliederung in Absätze kommt kaum bis gar nicht vor.

Mittel der szenischen Vergegenwärtigung, wie etwa direkte Rede und andere sprachliche narrative Elemente, wie zum Beispiel Metaphern oder Vergleiche, finden sich nicht. Auch evaluativ-emotionale Qualifizierungen der Ereignisse durch den Erzähler oder durch Aktanten mit dem Ziel, die Leser nicht nur zu informieren, sondern auch spannend zu unterhalten, enthalten die Texte nicht. Die Tempuswahl ist teilweise unzutreffend oder es

treten nichtfunktionale Tempuswechsel auf. Der Wortschatz, soweit er nicht der Aufgabe entnommen werden konnte, ist schmal und teilweise unangemessen. Der Satzbau ist in der Regel wenig variabel und parataktisch.

In der Mehrheit der Texte häufen sich grammatische und orthografische Fehler. Zuweilen werden Sätze formal, also durch Satzschlusszeichen, gar nicht abgegrenzt. Oftmals wird der Satzbau nicht oder nur unzureichend beherrscht. Die Kommaschreibung weist erhebliche Mängel auf. Einzelne Wörter und Wendungen werden häufig inkorrekt gebraucht.

Kompetenzstufe II (412 bis 492 Punkte; Mindeststandard)

Schülerinnen und Schüler auf Kompetenzstufe II können die Aufgaben so bearbeiten, dass eine narrative Textstruktur ansatzweise erkennbar ist. Die Erzählung bleibt dabei in Teilen jedoch undeutlich, wenig entfaltet, mitunter aufreihend und wiederholend. Häufig fehlen zentrale Elemente oder sie sind unplausibel dargestellt. Ist eine prototypische „Höhepunkt"-Geschichte gefragt, sind konstitutive Elemente wie Komplikation und Auflösung oftmals nicht oder nur rudimentär erkennbar. Mitunter wird auf dieser Stufe aber eine Exposition (Einführung der Handelnden und ggf. des Schauplatzes und der Zeit), zuweilen auch eine Coda (erzähltypische Abschlusspassage) produziert. Enthält die Aufgabe umfangreiche inhaltliche Vorgaben, Hintergrundinformationen oder Mustervorlagen, orientieren sich die Schreibenden typischerweise an einem oder maximal zwei der dadurch vorgegebenen inhaltlichen Aspekte und greifen diese im eigenen Text auf. Die Schülerinnen und Schüler schreiben weitgehend detailarm. Nur selten wird auf Gedanken, Gefühle und Handlungsmotive Bezug genommen. Erzählrelevante beschreibende Anteile werden in der Regel nur rudimentär oder wenig zielführend ausgearbeitet.

Bisweilen wird eine Erste- oder Dritte-Person-Perspektive durchgehalten. Die Linearisierung von Ereignissen zu kohärenten Ereignisfolgen gelingt auf dieser Stufe jedoch nur selten. Handlungen bzw. Ereignisse stehen oft unverbunden nebeneinander. Die inhaltliche und sprachliche Unterscheidung von Exposition, Komplikation und Auflösung wird weitgehend nicht gemeistert. Zuweilen werden die zumeist kürzeren Texte aber formal und auch inhaltlich plausibel gegliedert.

Gelegentlich finden sich Mittel der szenischen Vergegenwärtigung, wie etwa direkte Rede, sowie andere sprachliche narrative Elemente, wie zum Beispiel Attributionen oder Vergleiche. Es kommen mitunter auch evaluativ-emotionale Qualifizierungen der Ereignisse durch den Erzähler oder durch Aktanten vor. Bisweilen ist das Erzähltempus angemessen, häufig kommt es aber auch zu nicht-funktionalen Tempuswechseln. Der Wortschatz, soweit er nicht der Aufgabe entnommen werden konnte, ist in der Regel schmal und teilweise unangemessen, der Satzbau wenig variabel und parataktisch.

In den Schülertexten häufen sich bisweilen grammatische und orthografische Fehler. Gelegentlich werden die Sätze formal, also durch Satzschlusszeichen, nicht abgegrenzt. Der Satzbau wird oftmals nicht durchgängig beherrscht, die Kommaschreibung gelingt zuweilen auch in einfacheren Fällen nicht. Einzelne Wörter und Wendungen werden teilweise inkorrekt gebraucht.

Kompetenzstufe III (493 bis 573 Punkte; Regelstandard)

Auf Kompetenzstufe III bearbeiten die Schülerinnen und Schüler die Aufgabenstellung insofern zufriedenstellend, als dass sie in der Regel deutlich erkennbar narrative Texte produzieren. Ist eine prototypische „Höhepunkt"-Geschichte gefragt, sind häufig konstitutive Elemente wie Komplikation und Auflösung erkennbar. Auf dieser Stufe werden oft auch eine Exposition (Einführung der Handelnden und ggf. des Schauplatzes und der Zeit) und eine Coda (erzähltypische Abschlusspassage) produziert. Enthält die Aufgabe umfangreiche inhaltliche Vorgaben, Hintergrundinformationen oder Mustervorlagen, orientieren sich die Schreibenden typischerweise an wenigen der dadurch vorgegebenen inhaltlichen Aspekte und greifen diese im eigenen Text auf. Die Schülerinnen und Schüler füllen einige wenige inhaltliche Aspekte mit Details. Zuweilen wird im Text auf Gedanken, Gefühle und Handlungsmotive eingegangen. Einige erzählrelevante beschreibende Anteile werden in der Regel hinreichend ausgearbeitet und zielführend umgesetzt.

Die Schülerinnen und Schüler produzieren in der Regel Texte mittlerer Länge und halten dabei weitgehend eine Erste- oder Dritte-Person-Perspektive durch. Gelegentlich gelingt die Linearisierung von Ereignissen zu kohärenten Ereignisfolgen, häufig werden Handlungs- bzw. Ereignisfolgen aber noch nicht kohärent dargestellt. Die Texte sind oftmals grafisch-formal und auch inhaltlich plausibel gegliedert. Die inhaltliche und sprachliche Unterscheidung von Exposition, Komplikation und Auflösung gelingt jedoch nur manchmal.

Schülerinnen und Schüler auf dieser Stufe versuchen weitestgehend, den Leser möglichst spannend zu unterhalten, z. B. indem sie an manchen Stellen Mittel der szenischen Vergegenwärtigung einsetzen oder evaluativ-emotionale Qualifizierungen der Ereignisse vornehmen. Das Tempus ist dabei in der Regel funktional angemessen, nicht-funktionale Tempuswechsel kommen selten vor. Der Wortschatz, soweit er nicht der Aufgabe entnommen werden konnte, ist zumeist schmal, aber häufig angemessen. Der Satzbau ist auf dieser Stufe oftmals noch wenig variabel.

Schülerinnen und Schüler auf dieser Stufe zeigen weitestgehend Sicherheit im Satzbau und in der Beachtung von Satzgrenzen. Fehler in der Grammatik, in der Orthografie und in der Kommaschreibung treten bisweilen auf, beeinträchtigen das Verständnis jedoch in der Regel nicht. Mehrheitlich werden Wörter und Wendungen korrekt gebraucht.

Kompetenzstufe IV (574 bis 654 Punkte, Regelstandard plus)

Schülerinnen und Schüler auf Kompetenzstufe IV können ihre Texte in der Regel so gestalten, dass ein narrativer Zusammenhang gegeben ist. Ist eine prototypische „Höhepunkt"-Geschichte gefragt, sind die meisten konstitutiven Elemente, d. h. Exposition (Einführung der Handelnden und ggf. des Schauplatzes und der Zeit), Komplikation, Auflösung und Coda (erzähltypische Abschlusspassage) realisiert; zuweilen fehlt jedoch die Darstellung der Auflösung einer Komplikation. Enthält die Aufgabe umfangreiche inhaltliche Vorgaben, Hintergrundinformationen oder Mustervorlagen, orientieren sich die Schreibenden typischerweise an einigen bis mehreren der dadurch vorgegebenen inhaltlichen Aspekte und greifen diese im eigenen Text auf. Häufig werden einzelne Etappen, Situationen oder Elemente detailliert ausgeführt. Es finden sich mehrheitlich Referenzen auf Gedanken, Gefühle und Handlungsmotive. Erzählrelevante beschreibende Anteile werden weitgehend ausgearbeitet und zielführend umgesetzt.

In den eher längeren Texten wird die Erste- oder Dritte-Person-Perspektive größtenteils aufrechterhalten. In der Mehrheit der Fälle gelingt den Schülerinnen und Schülern die Linearisierung von Ereignissen zu kohärenten Ereignisfolgen, semantische Lücken kommen kaum vor. Konstitutive Elemente wie Exposition, Komplikation und Auflösung (falls gegeben) werden in der Regel nicht nur inhaltlich, sondern auch sprachlich angemessen markiert. Eine grafisch-formale sowie inhaltlich plausible Gliederung ist meist gegeben.

Schülerinnen und Schüler auf dieser Stufe versuchen durchweg, den Leser möglichst spannend zu unterhalten, z. B. indem sie Mittel der szenischen Vergegenwärtigung verwenden oder evaluativ-emotionale Qualifizierungen der Ereignisse vornehmen. Auch sprachliche Mittel wie Metaphern werden bisweilen eingesetzt. Das Tempus ist in der Regel angemessen. Der Wortschatz, der über das in der Aufgabe vorgegebene Vokabular hinausgeht, ist zuweilen breit und größtenteils angemessen. Der Satzbau ist auf dieser Stufe oft variabel und komplex.

Schülerinnen und Schülern auf dieser Stufe unterlaufen kaum Fehler in Orthografie und Grammatik, sie zeigen Sicherheit im Satzbau und in der Beachtung von Satzgrenzen. Auch die Kommaschreibung, insbesondere bei Aufzählungen und zur Abtrennung von Nebensätzen, wird in der Regel gemeistert. Wörter und Wendungen werden weitgehend korrekt gebraucht.

Kompetenzstufe V (ab 655 Punkten; Optimalstandard)

Schülerinnen und Schüler auf Kompetenzstufe V meistern den Großteil der Anforderungen, die mit dem Schreiben narrativer Texte verbunden sind. Ist eine prototypische „Höhepunkt"-Geschichte gefragt, sind in der Regel alle konstitutiven Elemente, d. h. Exposition (Einführung der Handelnden und ggf. des Schauplatzes und der Zeit), Komplikation, Auflösung und Coda (erzähltypische Abschlusspassage) realisiert; nicht-funktionale Wiederholungen oder Redundanzen kommen nicht vor. Enthält die Aufgabe umfangreiche inhaltliche Vorgaben, Hintergrundinformationen oder Mustervorlagen, orientieren sich die Schreibenden typischerweise an mehreren bis vielen der dadurch vorgegebenen inhaltlichen Aspekte und greifen diese im eigenen Text auf. Einzelne Etappen, Situationen oder Elemente werden weitgehend detailliert ausgeführt; Gedanken, Gefühle und Handlungsmotive werden fast durchweg thematisiert. Es gelingt den Schülerinnen und Schülern auf dieser Stufe, die meisten erzählrelevanten beschreibenden Anteile hinreichend auszuarbeiten und zielführend umzusetzen.

In den typischerweise längeren Texten wird die Erste- oder Dritte-Person-Perspektive nahezu immer durchgehalten. Meistens gelingt die Linearisierung von Ereignissen zu kohärenten Ereignisfolgen, semantische Lücken finden sich kaum. Die komplexe Aufgabe, Exposition, Komplikation und Auflösung nicht nur inhaltlich, sondern auch sprachlich angemessen zu gestalten, wird überwiegend gemeistert. Eine grafisch-formale sowie inhaltlich plausible Gliederung liegt weitestgehend vor.

Den Schülerinnen und Schülern geht es erkennbar darum, den Leser möglichst spannend zu unterhalten. So wird das Geschehen szenisch vergegenwärtigt und es finden sich evaluativ-emotionale Qualifizierungen der Ereignisse. Sprachliche Mittel, wie etwa Metaphern, werden häufig eingesetzt. Das Tempus ist durchgängig angemessen. Der Wortschatz, soweit er nicht der Aufgabe entnommen werden konnte, ist häufig breit, der Satzbau mehrheitlich variabel.

Schülerinnen und Schüler auf dieser Stufe zeigen Sicherheit im Satzbau und in der Beachtung von Satzgrenzen. Orthografische und grammatische Fehler treten nur vereinzelt auf. Die Kommaschreibung gelingt häufig auch in schwierigeren Fällen. Wörter und Wendungen werden weitestgehend korrekt gebraucht.

Tabelle A.4.2.1: ***Schülerverteilung in der Normierungsstudie – Kreuztabelle: Klassenstufe × Schulform inkl. Chi-Quadrat-Test auf Unabhängigkeit.***

		Schulform					
		Hauptschule	MBM	Integr. Gesamtschule	Realschule	Gymnasium	Gesamt
Klassenstufe	9	262	94	203	528	755	1842
	10	122	69	134	345	484	1154
Gesamt		384	163	337	873	1239	2996

$\chi^2 = 9.1$; $p = .058$

Tabelle A.4.2.2: ***Schülerverteilung in der Normierungsstudie – Kreuztabelle: Klassenstufe × Geschlecht inkl. Chi-Quadrat-Test auf Unabhängigkeit.***

		Geschlecht		
		männlich	weiblich	Gesamt
Klassenstufe	9	908	934	1842
	10	563	591	1154
Gesamt		1471	1525	2996

$\chi^2 = 0.07$; $p = .787$

Tabelle A.4.2.3: ***Schülerverteilung in der Normierungsstudie – Kreuztabelle: Klassenstufe × Sprachhintergrund inkl. Chi-Quadrat-Test auf Unabhängigkeit.***

		Deutsch als Herkunftssprache		
		nein	ja	Gesamt
Klassenstufe	9	232	1609	1841
	10	142	1012	1154
Gesamt		374	2621	2995

$\chi^2 = 0.06$; $p = .811$

Tabelle A.4.2.4: ***Schülerverteilung in der Normierungsstudie – Kreuztabelle: Geschlecht × Schulform inkl. Chi-Quadrat-Test auf Unabhängigkeit.***

		Schulform					
		Hauptschule	MBM	Integr. Gesamtschule	Realschule	Gymnasium	Gesamt
Geschlecht	männlich	207	91	185	386	602	1471
	weiblich	177	72	152	487	637	1525
Gesamt		384	163	337	873	1239	2996

$\chi^2 = 19.5$; $p = .001$

Tabelle A.4.2.5: ***Schülerverteilung in der Normierungsstudie – Kreuztabelle: Geschlecht × Sprachhintergrund inkl. Chi-Quadrat-Test auf Unabhängigkeit.***

		Deutsch als Herkunftssprache		
		nein	ja	Gesamt
Geschlecht	männlich	190	1281	1471
	weiblich	184	1340	1524
Gesamt		374	2621	2995

$\chi^2 = 1.5$; $p = .484$

Tabelle A.4.2.6: ***Schülerverteilung in der Normierungsstudie – Kreuztabelle: Schulform × Sprachhintergrund inkl. Chi-Quadrat-Test auf Unabhängigkeit.***

		Deutsch als Herkunftssprache		
		nein	ja	Gesamt
Schulform	Hauptschule	92	292	384
	MBG	12	151	163
	Integrierte Gesamtschule	72	264	336
	Realschule	115	758	873
	Gymnasium	83	1156	1239
Gesamt		374	2621	2995

$\chi^2 = 113.1$; $p < .001$

Tabelle A.7.7.1: ***Ergebnisse der Zwei-Ebenen-Moderatoranalysen unter Einbeziehung des Faktors „mittlere Häufigkeitsklasse“ und eines weiteren Faktors (außer „Wörter pro Satz“).***

γ_{00}	γ_{10}	γ_{11}	γ_{12}	var(r)	$var(u_0)$	$var(u_1)$	$\Delta var(u_1)/ var(u_{1(Nullmodell)})$
γ_{12}: Zeichenzahl							
-0.005	0.508***	0.274*	-0.0001	0.725	0.017***	0.0015	0.88
γ_{12}: Silbenzahl							
-0.005	0.508***	0.274*	-0.0001	0.725	0.017***	0.0015	0.88
γ_{12}: Wortzahl							
-0.005	0.508***	0.266*	-0.0001	0.725	0.017***	0.0019	0.85
γ_{12}: Satzzahl							
-0.005	0.507***	0.247*	-0.0001	0.725	0.017***	0.0015	0.88
γ_{12}: Zeichen pro Wort							
-0.005	0.508***	0.333**	-0.0991	0.725	0.017***	0.0008	0.94
γ_{12}: Silben pro Wort							
-0.005	0.508***	0.304**	-0.2460	0.725	0.017***	0.0002	0.98
γ_{12}: Anzahl seltene Wörter°							
-0.005	0.508***	0.377*	-0.0085	0.725	0.017***	0.0010	0.92
γ_{12}: Anteil seltene Wörter°							
-0.005	0.508***	0.328*	-0.0166	0.725	0.017***	0.0008	0.94
γ_{12}: LIX							
-0.005	0.507***	0.284*	-0.0015	0.725	0.017***	0.0011	0.91
γ_{12}: Flesch							
-0.005	0.508***	0.314*	0.0021	0.725	0.017***	0.0005	0.96

γ_{11}: mittlere Häufigkeitsklasse * $p < .05$ ** $p < .01$ *** $p < .001$

° aufgrund wechselseitiger Abhängigkeiten der Faktoren, nur bedingt interpretierbar

Tabelle A.7.7.2: ***Ergebnisse der Zwei-Ebenen-Moderatoranalysen unter Einbeziehung des Faktors „Wörter pro Satz“ und eines weiteren Faktors (außer „mittlere Häufigkeitsklasse“).***

γ_{00}	γ_{10}	γ_{11}	γ_{12}	var(r)	$var(u_0)$	$var(u_1)$	$\Delta var(u_1)/ var(u_{1(Nullmodell)})$
γ_{12}: Zeichenzahl							
-0.005	0.506***	0.025*	-0.0001	0.725	0.017***	0.0017	0.86
γ_{12}: Silbenzahl							
-0.005	0.506***	0.025*	0.0245	0.725	0.017***	0.0017	0.86
γ_{12}: Wortzahl							
-0.005	0.506***	0.024*	-0.0001	0.725	0.017***	0.0018	0.85
γ_{12}: Satzzahl							
-0.005	0.506***	0.024*	-0.0001	0.725	0.017***	0.0018	0.85
γ_{12}: Zeichen pro Wort							
-0.005	0.506***	0.038***	-0.1982	0.725	0.017***	0.0010	0.92
γ_{12}: Silben pro Wort							
-0.005	0.506***	0.033**	-0.4548	0.725	0.017***	0.0002	0.98
γ_{12}: Anzahl seltene Wörter							
-0.005	0.506***	0.027*	-0.0022	0.725	0.017***	0.0016	0.87
γ_{12}: Anteil seltene Wörter							
-0.005	0.506***	0.029*	0.0293	0.725	0.017***	0.0009	0.93
γ_{12}: LIX°							
-0.005	0.506***	0.031*	-0.0042	0.725	0.017***	0.0004	0.96
γ_{12}: Flesch°							
-0.005	0.507***	0.037**	0.0054	0.725	0.017***	0.0002	0.98

γ_{11}: Wörter pro Satz * $p < .05$ ** $p < .01$ *** $p < .001$

° aufgrund wechselseitiger Abhängigkeiten der Faktoren, nur bedingt interpretierbar

Tabelle A.7.7.3: ***Ergebnisse der Zwei-Ebenen-Moderatoranalysen unter Einbeziehung der Faktoren „mittlere Häufigkeitsklasse“, „Wörter pro Satz“ und eines weiteren Faktors.***

γ_{00}	γ_{10}	γ_{11}	γ_{12}	γ_{13}	var(r)	$var(u_0)$	$var(u_1)$	$\Delta var(u_1)/ var(u_{1(Nullmodell)})$
$\gamma_{13:}$ Zeichenzahl								
-0.005	0.508***	0.229	0.0047	-0.0001	0.725	0.017***	0.0017	0.86
$\gamma_{13:}$ Silbenzahl								
-0.005	0.508***	0.227	0.0048	-0.0001	0.725	0.017***	0.0017	0.86
$\gamma_{13:}$ Wortzahl								
-0.005	0.508***	0.223	0.0047	-0.0001	0.725	0.017***	0.0019	0.85
$\gamma_{13:}$ Satzzahl								
-0.005	0.507***	0.224	0.0026	-0.0008	0.725	0.017***	0.0019	0.85
$\gamma_{13:}$ Zeichen pro Wort								
-0.005	0.508***	0.165	0.0023	-0.1761	0.725	0.017***	0.0005	0.96
$\gamma_{13:}$ Silben pro Wort								
-0.005	0.508***	0.149	0.0186	-0.4026	0.725	0.017***	0.0001	0.99
$\gamma_{13:}$ Anzahl seltene Wörter°								
-0.005	0.508***	0.301	0.0079	-0.0079	0.725	0.017***	0.0009	0.93
$\gamma_{13:}$ Anteil seltene Wörter°								
-0.005	0.508***	0.234	0.0104	-0.0179	0.725	0.017***	0.0007	0.94
$\gamma_{13:}$ LIX°								
-0.005	0.507***	0.163	0.0149	-0.0032	0.725	0.017***	0.0007	0.94
$\gamma_{13:}$ Flesch°								
-0.005	0.508***	0.153	0.0231	0.0048	0.725	0.017***	0.0001	0.99

γ_{11}: mittlere Häufigkeitsklasse γ_{12}: Wörter pro Satz * $p < .05$ ** $p < .01$ *** $p < .001$

° aufgrund wechselseitiger Abhängigkeiten der Faktoren, nur bedingt interpretierbar